

MULTIRESOLUTION DIGITAL SOIL MAPPING OF PERMAFROST SOILS USING A  
RANDOM FOREST CLASSIFIER: AN INVESTIGATION ALONG THE DALTON HIGHWAY  
CORRIDOR, ALASKA

By

Joshua D Paul, B.A.

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in

Permafrost Soils and Digital Soil Mapping: Interdisciplinary Program

University of Alaska Fairbanks

December 2018

APPROVED:

Chien-Lu Ping, Committee Co-Chair  
Anupma Prakash, Committee Co-Chair  
Jordi Cristobal Rossello, Committee Member  
Zamir Libohova, Committee Member  
Joshua Greenberg, Chair  
*Department of Natural Resource Management*  
David Valentine, Director of Academic Programs  
*School of Natural Resources and Extension*  
Michael Castellini, *Dean of Graduate School*

## **Abstract**

In order to complete soil inventories in the remote permafrost zones of Alaska, there is a need to develop efficient digital soil mapping tools that can be applied over large areas using a minimum of ground truth data.

This investigation first used a random forest classifier to test combinations of environmental input data at multiple resolutions (10m, 30m, and 100m). Five tiers of soil taxonomic units were predicted: Order, Suborder, Great Group, “Series Concept”, and Particle Size Class. Model outputs are compared quantitatively via estimated out-of-bag accuracy, and qualitatively via visual inspection by soil scientists. Estimated out-of-bag accuracy ranged from ~45% to ~75%, with results improving when fewer classes were modeled. Model runs at 10m and 30m resolution performed comparably, with 100m resolution performing ~5-10% worse in most cases. Increasing the number of trees used, including categorical environmental input data (e.g. landforms), and replacement of environmental covariates with principal component analysis (PCA) bands did not significantly improve model performance.

The random forest classifier was then used in a digital soil mapping pilot study along the Dalton Highway in northern Alaska. Parameters suggested in the initial study were used to predict multiple soil taxonomic classes from a basic collection of environmental covariates generated using high resolution (10m) satellite images and sparsely sampled pedon data. Covariates included maximum curvature, multiresolution valley bottom flatness, normalized height, potential incoming solar radiation, slope, terrain ruggedness index, and modified soil and vegetation index. Five tiers of soil taxonomic units were predicted: Order, Suborder, Great Group, “Series Concept”, and Particle Size Class. Model outputs are compared quantitatively via estimated out-of-bag accuracy. Estimated out-of-bag accuracy ranged from ~45% to ~75%, with results improving when fewer classes were modeled.

We suggest future research into optimized sampling to ensure an adequate distribution of samples across the feature space, and the incorporation of expert knowledge into accuracy assessments. Overall, digital soil mapping with random forest classifiers appears to be a promising method for completing the soil survey of Alaska.

Title Page .....	i
Abstract .....	ii
Table of Contents .....	iv
List of Tables .....	vi
List of Figures .....	vii
CHAPTER 1: INTRODUCTION .....	1
1.1 Background .....	1
1.2 Status of Soil Mapping in Alaska .....	1
1.3 Digital Soil Mapping Using A Random Forest Classifier .....	2
1.4 Study Area .....	4
1.5 References .....	5
CHAPTER 2: MULTIREOLUTION DIGITAL SOIL MAPPING OF PERMAFROST SOILS USING A RANDOM FOREST CLASSIFIER .....	10
2.1 Abstract .....	10
2.2 Introduction .....	11
2.2.1 Status of Soil Mapping in Alaska .....	11
2.2.2 Study Area .....	13
2.3 Data Acquisition and Preprocessing .....	15
2.3.1 Soils Data .....	15
2.3.2 Multispectral Imagery .....	16
2.3.3 Elevation Data .....	17
2.3.4 Geologic Data .....	17
2.3.5 Geomorphons .....	17
2.3.6 Terrain Attributes .....	18
2.3.7 Layer Stacking, Principal Component Analysis, and Band Statistics .....	19
2.4 Methods .....	20
2.4.1 Random Forest Classification .....	20
2.4.2 Accuracy Assessment .....	21
2.5 Results .....	22
2.5.1 Estimated OOB Accuracy by Taxonomic Class and Resolution .....	23
2.5.2 Visual Inspection of Map Results .....	24
2.6 Discussion .....	24
2.6.1 Stacks .....	24
2.6.2 Resolution .....	25
2.6.3 Classification Schema .....	26
2.6.4 Trees .....	27

2.6.5 Pedon Data .....	27
2.6.6 Map Quality .....	30
2.6.7 Recommendations for Post-Processing.....	34
2.7 Conclusions.....	36
2.8 Acknowledgments.....	37
2.9 References.....	37
2.10 Appendix A .....	63
2.11 Appendix B.....	64
<b>CHAPTER 3: DIGITAL SOIL MAPPING OF PERMAFROST SOILS ALONG THE DALTON HIGHWAY CORRIDOR, ALASKA: A PILOT STUDY USING A RANDOM FOREST CLASSIFIER .....</b>	<b>65</b>
3.1 Abstract .....	65
3.2 Introduction.....	66
3.2.1 Study Area.....	66
3.3 Data Acquisition and Preprocessing .....	67
3.3.1 Soils Data .....	67
3.3.2 Environmental Covariates.....	68
3.4 Methods.....	68
3.4.1 Random Forest Classification .....	68
3.4.2 Accuracy Assessment.....	69
3.5 Results and Discussion.....	70
3.5.1 Estimated OOB Accuracy and Kappa by Taxonomic Class .....	70
3.5.2 Importance Table.....	70
3.6 Discussion .....	70
3.6.1 Importance of Environmental Covariates .....	70
3.6.2 Covariate Resolution.....	72
3.6.3 Classification Schema.....	73
3.6.4 Pedon Data .....	73
3.6.5 Quantitative Accuracy.....	75
3.6.6 Recommendations for Post-Processing.....	75
3.7 Conclusions.....	77
3.8 Acknowledgments.....	78
3.9 References.....	78
<b>CHAPTER 4: CONCLUSION.....</b>	<b>88</b>

## List of Tables

Page

<b>Table 2.1</b> Pedon locations and classifications used for modeling. ....	51
<b>Table 2.2</b> Environmental covariate layers included in each raster stack. ....	56
<b>Table 2.3</b> Results of visual evaluations by three soil scientists. Each mapset received a maximum of 60 points from each individual evaluation, and the results are sorted by evaluator mean rating. ....	57
<b>Table 3.1</b> Estimated OOB Accuracy and Kappa by taxonomic level modeled. ....	86
<b>Table 3.2</b> Mean Decrease in Accuracy (MDA) values for each environmental covariate and taxonomic level modeled. ....	87

## List of Figures

## Page

<b>Figure 2.1</b> Status map showing SSURGO soil surveys, ABR soil surveys, and unsurveyed areas in Alaska. The Dalton Highway corridor DSM research area is also highlighted .....	58
<b>Figure 2.2</b> Detailed map of Dalton Highway corridor DSM research area, with landscape photos and descriptions of major physiographic regions .....	59
<b>Figure 2.3.</b> Soil pit photos and descriptions of series concepts .....	60
<b>Figure 2.4</b> Boxplots of model accuracy.....	62

## **CHAPTER 1: INTRODUCTION**

### **1.1 Background**

Federally-supported soil survey in the United States of America (USA) began in the late 1890s, and included lands in the territory of Alaska by the 1910s (Baker, 1963; Bennett and Rice, 1915; Bennett, 1919; Soil Survey Staff, 1993). Early soil surveys in Alaska focused on small areas of potential agricultural land and were often conducted as part of exploratory geographic mapping, geologic and mining surveys, and the establishment of agricultural experiment stations (Mitchell, 1998; Bennett and Rice, 1915; Bennett, 1919; Sherwood, 1965). While the contiguous USA currently has nearly complete soil survey coverage at scales of 1:63,630 or finer, a large percentage of the state of Alaska remains unsurveyed at scales finer than 1:500,000 (Soil Survey Staff, 2018). After over a century of soil survey in the state, there is still much work to be done. In order to complete the soil survey at scales appropriate for land management decisions, new approaches are required to improve efficiency and counter shrinking budgets.

### **1.2 Status of Soil Mapping in Alaska**

Current mapping efforts in Alaska are produced and delivered digitally at much finer scales and are intended for a wider range of users when compared to these legacy surveys. In general, the current state of soils mapping in Alaska can generally be divided into two distinct groups: soil property maps and soil classification maps.

Digital mapping of continuous or classified soil properties in Alaska has grown along with worldwide interest in climate change. Climate modeling in the remote, sparsely sampled Arctic and subarctic regions of Alaska has demanded rapid production of datasets representing soil information. Thus far, the parameters of soil organic carbon (SOC) stocks and active layer thickness (ALT) have been mapped most extensively using remote sensing techniques (Deluca and Boisvenue, 2012; Hugelius, 2012; Hugelius et al., 2014; Jafarov et al., 2012; Mishra and Riley,



2012, 2014; Panda et al., 2010; Panda, 2014; Pastick et al., 2013, 2014; Ping et al., 2008a). Both SOC and ALT are considered to be critical baseline data in modeling climate change and the impact of increasing temperatures on Arctic and subarctic infrastructure.

Mapping of soil classifications in Alaska is an ongoing effort, far behind the status of soil mapping in the conterminous United States. Soil classification maps produced by the United States Department of Agriculture, Natural Resources Conservation Service (USDA-NRCS) are designed to map soil types explicitly and sometimes include associated detailed vegetation information (depending on vintage). USDA-NRCS has completed statewide soils mapping at the Digital General Soil Map of the United States standard (STATSGO2, 1:63,360 and coarser scales), and has mapped approximately 20% of the state at the finer-resolution Soil Survey Geographic Database standard (SSURGO, 1:63,360 and finer scales) (National Soil Survey Staff, 2017; Soil Survey Staff, 2018) (Figure 2.1). Soil maps produced by the private sector are best thought of as vegetation and landform maps, with soils information associated with each mapunit. Private sector "soil landscapes" mapping in Alaska primarily includes work by ABR Inc., mapping entire regions of Alaska as well as smaller inventory projects for the National Park Service and lands leased by the oil and gas industry (Jorgenson, et al., 2003, 2009; Wells et al., 2013). Mapping scale, map unit design, sampling schemes, and soil profile information vary widely between existing soil class maps, as does the degree to which vegetation and ecological data drive the mapping process and final map products.

### **1.3 Digital Soil Mapping Using A Random Forest Classifier**

This research focuses on random forest classification - a digital soil mapping methodology not yet used in the state of Alaska - and applies this classification method to a remote, sparsely sampled area along the Dalton Highway corridor. The thesis includes this introduction, followed by two major chapters that are stand-alone manuscripts to be submitted for journal publication, and a final conclusion.

The first chapter is an investigation into the effects of different input data layers used in automated digital soil mapping, comparing models built at various resolutions, levels of soil taxonomy, and number of trees used in the random forest classification algorithm. The goal is to develop a digital soil mapping methodology for the United States Department of Agriculture, Natural Resource Conservation Service's initial soils mapping of Alaska. Using a random forest classification, the overall accuracy of map results is assessed using both point-based ground truth data and visual inspection by soil experts with experience in the study area. Using these assessments, this research aims to a) determine the most appropriate modeling resolution; b) determine the most appropriate parameters to use in modeling; and c) determine the most appropriate number of trees to build in the random forest algorithm. A variety of categorical and continuous environmental covariates are tested, and the entire analysis is performed at three different resolutions (10m, 30m, and 100m) to compare accuracies between each. The number of trees built in each random forest modeling run is also varied for each resolution and combination of layers. In total, 432 models are compared.

The second chapter uses the conclusions of the previous investigation to create a digital soil map of the Dalton Highway corridor using a mixture of legacy pedon data and data gathered specifically for this research. This pilot study is an attempt to approach the Soil Survey Geographic Database (SSURGO) standard of mapping taxonomic soil classes at scales of less than or equal to 1:63,360 (National Soil Survey Staff, 2017) by applying a random forest classification method. Using common environmental covariates at high resolution (10m), this study presents a baseline modeling accuracy that can be expected when using freely available data layers and limited distribution of sampling points. The overall accuracy of map results is assessed using point-based ground truth data. With a sampling density of 106 direct soil observations for a mapping area of 12,088 km<sup>2</sup> (1 observation per ~114 km<sup>2</sup>), this research should serve as an appropriate pilot study for digital soil mapping in remote areas of Alaska with limited pedon data.

## 1.4 Study Area

This research focuses on an approximately 10 miles corridor on either side of the Dalton Highway, extending from Atigun Pass in the Brooks Range to the terminus of the highway at Deadhorse, Alaska (Figure 2.2). Land ownership in this area is primarily federal (United States Department of the Interior - Bureau of Land Management) and State of Alaska. The majority of the study area is within the zone of continuous permafrost, where over 90% of the earth's surface contains materials at or below 0 C for two or more consecutive years (Washburn, 1973). However, from the limited depth perspective of soil classification (approximately 2 m) there are many instances of unfrozen soils that occur in this region.

At least three major physiographic regions are included within the study area boundary. In the south, the highway passes over the Brooks Range Mountains. Though the route through Atigun Pass is relatively low elevation (~1415m), the entirety of the mountainous terrain in the study area is in the alpine life zone due to its high latitude. The northern limit of tree growth occurs on the southern slopes of the range, with arctic tundra vegetation occurring north of the range (Gallant et al., 1995). The geology of the Brooks Range in this area is dominantly sedimentary rocks of marine and deltaic origin, with the oldest formations occurring in the vicinity of Atigun Pass and the youngest formations exposed further north in the Arctic Foothills. (Huryn and Hobbie, 2012; Wilson et al., 2015). Owing to late Pleistocene glaciation, the slopes of the Brooks Range are very rugged and have been predicted to have very little soil development (Gallant et al., 1995; Huryn and Hobbie, 2012). The Sagavinirktok River and all its tributaries have their headwaters in the Brooks Range and flow north through the study area.

North of the range the highway passes through the Arctic Foothills. This region includes rolling moraine, kame and kettle complexes, floodplain and terrace complexes along major rivers, and occasional bedrock outcroppings (Hamilton, 2003). The oldest glaciated surfaces have rounded slope shapes, variable depths of loess over the glacial till, and generally lack rock fragments at the surface (Walker et al., 2014). Younger glaciated surfaces are more variable with regards to slope

shape, lack any loess cap, and have substantial exposed surface rock fragments (Walker et al., 2014).

Between the Arctic Foothills and the Arctic Ocean, the highway passes through and terminates within the Arctic Coastal Plain. With very low relief, this area is dominated by fluvial sediments mantled with loess material that has been cryoturbated into polygonal patterned ground (Ping et al., 1998, 2008b, 2013). Numerous oblong, oriented thaw lakes occur throughout the plain and can be seen in various stages of draining or filling (Hinkel et al., 2003). Though the terrain is flat when viewed at coarse scale, the combination of low-centered and high-centered polygonal patterned ground features offer substantial micro-relief and potentially contain many soil components (Ping et al., 2013). However, these features are generally only resolvable at very fine scales and do not appear on most medium resolution imagery or elevation data. For this reason, this region presents unique challenges to digital soil mapping that may not be present in warmer climates.

## **1.5 References**

- Baker, G. L. (1963). Century of service: the first 100 years of the United States Department of Agriculture. Centennial Committee, US Department of Agriculture. Washington, D.C.
- Bennett, H. H. (1919). Report on a reconnaissance of the soils, agriculture, and other resources of the Kenai Peninsula Region of Alaska. US Government Printing Office. Washington, D.C.
- Bennett, H. H., & Rice, T. D. (1915). Soil Reconnaissance in Alaska: With an Estimate of the Agricultural Possibilities. US Government Printing Office. Washington, D.C.
- Deluca, T. H., & Boisvenue, C. (2012). Boreal forest soil carbon: distribution, function and modelling. *Forestry*, 85(2), 161–184. <https://doi.org/10.1093/forestry/cps003>
- Gallant, A. L., Binnian, E. F., Omernik, J. M., & Shasby, M. B. (1995). Ecoregions of Alaska (USGS Numbered Series No. 1567) (p. 78). U.S. Geological Survey. Retrieved from <http://pubs.er.usgs.gov/publication/pp1567>

- Hamilton, T. D. (2003). Glacial geology of the Toolik Lake and upper Kuparuk River regions. Institute of Arctic Biology. University of Alaska Fairbanks.
- Hinkel, K. M., Eisner, W. R., Bockheim, J. G., Nelson, F. E., Peterson, K. M., & Dai, X. (2003). Spatial Extent, Age, and Carbon Stocks in Drained Thaw Lake Basins on the Barrow Peninsula, Alaska. *Arctic, Antarctic, and Alpine Research*, 35(3), 291–300.  
[https://doi.org/10.1657/1523-0430\(2003\)035\[0291:SEAACS\]2.0.CO;2](https://doi.org/10.1657/1523-0430(2003)035[0291:SEAACS]2.0.CO;2)
- Hugelius, G. (2012). Northern Circumpolar Soil Carbon Database. ECDS Environment Climate Data Sweden. <https://doi.org/10.5879/ecds/000000001>
- Hugelius, G., Strauss, J., Zubrzycki, S., Harden, J. W., Schuur, E. A. G., Ping, C.-L., ... Kuhry, P. (2014). Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps. *Biogeosciences*, 11(23), 6573–6593.  
<https://doi.org/10.5194/bg-11-6573-2014>
- Huryn, A. D., & Hobbie, J. E. (2012). Land of extremes: a natural history of the Arctic North Slope of Alaska. Fairbanks, Alaska: University of Alaska Press.
- Jafarov, E. E., Marchenko, S. S., & Romanovsky, V. E. (2012). Numerical modeling of permafrost dynamics in Alaska using a high spatial resolution dataset. *The Cryosphere*, 6(3), 613–624.  
<https://doi.org/10.5194/tc-6-613-2012>
- Jorgenson, M. T., Roth, J. E., Emers, M., Schlentner, S. F., Swanson, D. K., Pullman, E. R., & Stickney, A. A. (2003). An ecological land survey in the Northeast Planning Area of the National Petroleum Reserve–Alaska, 2002. ABR. Inc., Fairbanks, AK.
- Jorgenson, M. T., Roth, J. E., Miller, P. F., Macander, M., Duffy, M. S., Wells, A., ... Pullman, E. R. (2009). An ecological land survey and landcover map of the Arctic Network (Natural Resource Technical Report No. NPS/ARC/NRTR—2009/270). Fort Collins, Colorado: National Park Service. Retrieved from  
[http://science.nature.nps.gov/im/units/arcn/documents/documents/NPS\\_ARCN\\_NRTR-2009-270\\_EcologicalLandSurveyLandcoverMap.pdf](http://science.nature.nps.gov/im/units/arcn/documents/documents/NPS_ARCN_NRTR-2009-270_EcologicalLandSurveyLandcoverMap.pdf)

- Mishra, U., & Riley, W. J. (2012). Alaskan soil carbon stocks: spatial variability and dependence on environmental factors. *Biogeosciences*, 9(9), 3637–3645. <https://doi.org/10.5194/bg-9-3637-2012>
- Mishra, U., & Riley, W. J. (2014). Active-Layer Thickness across Alaska: Comparing Observation-Based Estimates with CMIP5 Earth System Model Predictions. *Soil Science Society of America Journal*, 78(3), 894. <https://doi.org/10.2136/sssaj2013.11.0484>
- Mitchell, G.A. (1998). Acting Director's Letter: 100 Years of Alaska Agriculture. *Agroborealis* 30 (1): 1. , Spring 1998. School of Agriculture and Land Resources Management Agricultural and Forestry Experiment Station, University of Alaska Fairbanks.
- National Soil Survey Staff. SSURGO/STATSGO2 Structural Metadata and Documentation; NRCS Soils. Retrieved November 5, 2017. [https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2\\_053631](https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053631)
- Panda, S. K. (2014). High-Resolution Permafrost Modeling in Denali National Park and Preserve (Natural Resource Technical Report No. NPS/CAKN/NRTR—2014/858). National Park Service. Fort Collins, Colorado. Retrieved from <https://irma.nps.gov/App/Reference/Profile/2208990>
- Panda, S. K., Prakash, A., Solie, D. N., Romanovsky, V. E., & Jorgenson, M. T. (2010). Remote sensing and field-based mapping of permafrost distribution along the Alaska Highway corridor, interior Alaska. *Permafrost and Periglacial Processes*, 21(3), 271–281. <https://doi.org/10.1002/ppp.686>
- Pastick, N. J., Jorgenson, M. T., Wylie, B. K., Minsley, B. J., Ji, L., Walvoord, M. A., ... Rose, J. R. (2013). Extending Airborne Electromagnetic Surveys for Regional Active Layer and Permafrost Mapping with Remote Sensing and Ancillary Data, Yukon Flats Ecoregion, Central Alaska: Remote Sensing and Mapping of Permafrost and Active-layer Thickness. *Permafrost and Periglacial Processes*, 24(3), 184–199. <https://doi.org/10.1002/ppp.1775>

- Pastick, N. J., Rigge, M., Wylie, B. K., Jorgenson, M. T., Rose, J. R., Johnson, K. D., & Ji, L. (2014). Distribution and landscape controls of organic layer thickness and carbon within the Alaskan Yukon River Basin. *Geoderma*, 230–231, 79–94.  
<https://doi.org/10.1016/j.geoderma.2014.04.008>
- Ping, C. L., Bockheim, J. G., Kimble, J. M., Michaelson, G. J., & Walker, D. A. (1998). Characteristics of cryogenic soils along a latitudinal transect in arctic Alaska. *Journal of Geophysical Research: Atmospheres*, 103(D22), 28917–28928.  
<https://doi.org/10.1029/98JD02024>
- Ping, C. L., Michaelson, G. J., Kimble, J. M., Romanovsky, V. E., Shur, Y. L., Swanson, D. K., & Walker, D. A. (2008b). Cryogenesis and soil formation along a bioclimate gradient in Arctic North America. *Journal of Geophysical Research*, 113(G3).  
<https://doi.org/10.1029/2008JG000744>
- Ping, C.-L., Clark, M. H., Kimble, J. M., Michaelson, G. J., Shur, Y., & Stiles, C. A. (2013). Sampling Protocols for Permafrost-Affected Soils. *Soil Horizons*, 54(1), 13.  
<https://doi.org/10.2136/sh12-09-0027>
- Ping, C.-L., Michaelson, G. J., Jorgenson, M. T., Kimble, J. M., Epstein, H., Romanovsky, V. E., & Walker, D. A. (2008a). High stocks of soil organic carbon in the North American Arctic region. *Nature Geoscience*, 1(9), 615–619. <https://doi.org/10.1038/ngeo284>
- Sherwood, M. B. (1965). *Exploration of Alaska, 1865-1900*. Yale University Press, New Haven.
- Soil Survey Staff (1993). *Soil Survey Manual*. Soil Conservation Service. U.S. Department of Agriculture Handbook 18.
- Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. <http://websoilsurvey.nrcs.usda.gov>. Accessed February 2, 2018.
- Walker, D. A., Hamilton, T. D., Maier, H. A., Munger, C. A., & Raynolds, M. K. (2014). Glacial History and Long-Term Ecology in the Toolik Lake Region. In J. E. Hobbie & G. W. Kling

(Eds.), *Alaska's Changing Arctic* (pp. 61–80). Oxford University Press.

<https://doi.org/10.1093/acprof:osobl/9780199860401.003.0003>

Washburn, A. L. (1973). *Periglacial Processes and Environments*. St. Martin's Press, New York.

Wells, A., Macander, M., Jorgenson, M. T., Christopherson, T., Baird, B., Trainor, E., & Martyn, P.

(2013, June). *Ecological Land Classification, Soil Landscape Mapping, and Near Infrared*

*(NIR) Spectroscopy of Soils in Lake Clark National Park and Preserve*. *Alaska Park*

*Science*, 12(1), 51–59.

Wilson, F.H., Hults, C.P., Mull, C.G, and Karl, S.M, comps. (2015). *Geologic map of Alaska*: U.S.

*Geological Survey Scientific Investigations Map 3340*, pamphlet 196 p., 2 sheets, scale

1:1,584,000. <http://dx.doi.org/10.3133/sim3340>



## CHAPTER 2: MULTIREOLUTION DIGITAL SOIL MAPPING OF PERMAFROST SOILS USING A RANDOM FOREST CLASSIFIER<sup>1</sup>

### 2.1 Abstract

In order to complete soil inventories in the remote permafrost zones of Alaska, there is a need to develop efficient digital soil mapping tools that can be applied over large areas using a minimum of ground truth data. This investigation uses a random forest classifier to test combinations of environmental input data at multiple resolutions (10m, 30m, and 100m). Five tiers of soil taxonomic units are predicted: Order, Suborder, Great Group, “Series Concept”<sup>2</sup>, and Particle Size Class. Model outputs are compared quantitatively via estimated out-of-bag accuracy, and qualitatively via visual inspection by soil scientists. Estimated out-of-bag accuracy ranged from ~45% to ~75%, with results improving when fewer classes were modeled. Model runs at 10m and 30m resolution performed comparably, with 100m resolution performing ~5-10% worse in most cases. Increasing the number of trees used, including categorical environmental input data (e.g. landforms), and replacement of environmental covariates with PCA principal component bands did not significantly improve model performance. We suggest future research into optimized sampling to ensure an adequate distribution of samples across the feature space, and the incorporation of expert knowledge into accuracy assessments. Overall, digital soil mapping with random forest classifiers appears to be a promising method for completing the soil survey of Alaska.

---

<sup>1</sup> Paul, J. et al., 2018, Multiresolution Digital Soil Mapping of Permafrost Soils Using a Random Forest Classifier, (prepared for submission to Journal of Permafrost and Periglacial Processes)

<sup>2</sup> Series Concepts are further defined in section 2.2.1 of this manuscript.

## **2.2 Introduction**

Centralized, federally-supported soil mapping began in the conterminous United States well before Alaska achieved statehood in 1959. However, basic reconnaissance-level soil surveying began in Alaska prior to 1959 as part of exploratory geographic mapping, geologic and mining surveys, and the establishment of agricultural experiment stations (Mitchell, 1998; Sherwood, 1965). Current mapping efforts in Alaska are produced and delivered digitally at much finer scales and are intended for a wider range of users when compared to these legacy surveys.

### **2.2.1 Status of Soil Mapping in Alaska**

The current state of soils mapping in Alaska can generally be divided into two distinct groups: soil property maps and soil classification maps.

Digital mapping of continuous or classified soil properties in Alaska has grown along with worldwide interest in climate change. Climate modeling in the remote, sparsely sampled Arctic and subarctic regions of Alaska has demanded rapid production of datasets representing soil information. In particular, the parameters of soil organic carbon (SOC) stocks and active layer thickness (ALT) have been mapped extensively using remote sensing techniques (Deluca and Boisvenue, 2012; Hugelius, 2012; Hugelius et al., 2014; Jafarov et al., 2012; Mishra and Riley, 2012, 2014; Panda et al., 2010; Panda, 2014; Pastick et al., 2013, 2014; Ping et al., 2008a). Both SOC and ALT are considered to be critical baseline data in modeling climate change and the impact of increasing temperatures on Arctic and subarctic infrastructure.

Mapping of soil classifications in Alaska is an ongoing effort, far behind the status of soil mapping in the contiguous United States. Soil classification maps produced by the United States Department of Agriculture, Natural Resources Conservation Service (USDA-NRCS) are designed to map soil types explicitly and sometimes include associated detailed vegetation information (depending on vintage). USDA-NRCS has completed statewide soils mapping at the Digital General Soil Map of the United States standard (STATSGO2, 1:63,360 and coarser scales), and has mapped approximately 20% of the state at the finer-resolution Soil Survey Geographic Database standard (SSURGO, 1:63,360 and finer scales) (National Soil Survey Staff, 2017; Soil Survey Staff 2018) (Figure 2.1). Soil maps produced by the private sector are best thought of as vegetation and landform maps, with soils information associated with each mapunit. Private sector "soil landscapes" mapping in Alaska primarily includes work by ABR Inc., mapping entire regions of Alaska as well as smaller inventory projects for the National Park Service and lands leased by the oil and gas industry (Jorgenson, et al., 2003, 2009; Wells et al., 2013). Mapping scale, map unit design, sampling schemes, and soil profile information vary widely between existing soil class maps, as does the degree to which vegetation and ecological data drive the mapping process and final map products.

With the goal of developing a digital soil mapping methodology for the United States Department of Agriculture, Natural Resource Conservation Service's initial soils mapping of Alaska, this research focuses on approaching the Soil Survey Geographic Database (SSURGO) standard of mapping taxonomic soil classes at scales of less than or equal to 1:63,360 (National Soil Survey Staff, 2017). Using a random forest classification, the overall accuracy of map results is assessed using both point-based ground truth data and visual inspection by soil experts with experience in the study area. Using these assessments, this research aims to a) determine the most appropriate modeling resolution; b) determine the most appropriate parameters to use in modeling; and c) determine the most appropriate number of trees to build in the random forest algorithm. A variety of categorical and continuous environmental covariates (described below) are tested, and the

entire analysis is performed at three different resolutions (10m, 30m, and 100m) to compare accuracies between each. The number of trees built in each random forest modeling run is also varied for each resolution and combination of layers. In total, 432 models are compared. With a sampling density of 106 direct soil observations for a mapping area of 12,088 km<sup>2</sup> (1 observation per ~114 km<sup>2</sup>), this research will serve as an appropriate pilot study for digital soil mapping in remote areas of Alaska with limited pedon data.

### **2.2.2 Study Area**

This research focuses on an approximately 10 miles corridor on either side of the Dalton Highway, extending from Atigun Pass in the Brooks Range to the terminus of the highway at Deadhorse, Alaska (Figure 2.2). Land ownership in this area is primarily United States Department of the Interior - Bureau of Land Management, and State of Alaska property. The majority of the study area is within the zone of continuous permafrost, where over 90% of the earth's surface contains materials at or below 0 °C for two or more consecutive years (Washburn, 1973). However, from the limited depth perspective of soil classification, approximately 2 m as required for Gelisols (Soil Survey Staff, 1999), there are many instances of non-Gelisols that occur in this region.

At least three major physiographic regions are included within the study area boundary: In the south, the highway passes over the Brooks Range Mountains. Though the route through Atigun Pass is relatively low elevation (~1415m), the entirety of the mountainous terrain in the study area is in the alpine life zone due to its high latitude. The northern limit of tree growth occurs on the southern slopes of the range, with arctic tundra vegetation occurring north of the range (Gallant et al., 1995). The geology of the Brooks Range in this area is dominantly sedimentary rocks of marine and deltaic origin, with the oldest formations occurring in the vicinity of Atigun Pass and the youngest formations exposed further north in the Arctic Foothills. (Huryn and Hobbie, 2012; Wilson et al., 2015). Owing to late Pleistocene glaciation, the slopes of the Brooks Range are very rugged and have been predicted to have very little soil development (Gallant et al., 1995; Huryn and

Hobbie, 2012; Soil Survey Staff, 2018). The Sagavanirktok river and all its tributaries have their headwaters in the Brooks Range and flow north through the study area.

North of the mountains the highway passes through the Arctic Foothills. This region includes rolling glaciated plains, kame and kettle complexes, floodplain and terrace complexes along major rivers, and occasional bedrock outcroppings (Hamilton, 2003). The oldest glaciated surfaces have rounded slope shapes, variable depths of loess over the glacial till, and generally lack rock fragments at the surface (Walker et al., 2014). Younger glaciated surfaces are more variable with regards to slope shape, lack any loess cap, and have substantial exposed surface fragments (Walker et al., 2014).

Between the Arctic Foothills and the Arctic Ocean, the highway passes through and terminates within the Arctic Coastal Plain. With very low relief, this area is dominated by fluvial sediments mantled with loess material in which cryoturbated soils formed along with polygonal patterned ground development (Ping et al., 1998, 2008b, 2013). Numerous oblong, oriented thaw lakes occur throughout the plain and can be seen in various stages of draining or filling (Hinkel et al., 2003). Though the terrain is flat when viewed at coarse scale, the combination of low-centered and high-centered polygonal patterned ground features offer substantial micro-relief and potentially contain many soil components (Ping et al., 2013). However, these features are generally only resolvable at very fine scales and do not appear on most medium resolution imagery or elevation data. For this reason, this region presents unique challenges to digital soil mapping that may not be present in warmer climates.

## **2.3 Data Acquisition and Preprocessing**

### **2.3.1 Soils Data**

Historical field pedon data, pedon descriptions from field work specific to this study, and remotely sensed points were aggregated into a point data file for use in model training and accuracy assessment. All pedons were described to NRCS standards and most can be classified to the subgroup or family level. The total number of pedons in the dataset was 106, with an additional 24 points classified indirectly via remote sensing (Table 2.1). These additional points were in miscellaneous areas such as water bodies, gravel bars, and rock outcrops where the regolith is typically not classified as a soil but components are traditionally still included in soil map units.

Point data attributes included multiple levels of soil taxonomic classifications ranging from soil order down to particle-size class (Soil Survey Staff, 1999). As soil series have not yet been created for this region, unique "series concept" names and their corresponding taxonomic units were created specifically for this research (Figure 2.3). In this mapping context, these series concept names function similarly to complexes or associations of soil series, but individual units are not as narrowly defined with regards to Soil Taxonomy. Each pedon was correlated to the series concept of closest fit based on landform, parent material, and basic taxonomic unit. As a result, pedon data labeled with a unique series concept name may have multiple taxonomic classifications associated with it. The "series concept" can be considered as a cluster of closely related soil types with similar interpretive properties.

Some points in the dataset are classified with multiple, concatenated series concept names and miscellaneous area types. These points represent complexes of soil and miscellaneous areas that occur at the sub-pixel scale (for example, shallow soils co-occurring with rock outcrops on a mountain summit). These points occur mostly in the mountainous regions of the mapping areas where "pure" pixels of a given soil type are uncommon.

### 2.3.2 Multispectral Imagery

In this study, Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) multispectral imagery was used to capture the most recent state of vegetation in the study area with the intention of creating vegetation indices for modeling. Over 90% of the study area had cloud- and snow-free images available from July 2016. Non-thermal bands from two scenes (LC80730112016206LGN00 and LC81472322016197LGN00) were converted to reflectance values following the true "Top of Atmosphere" correction procedure detailed in the Landsat 8 Data Users Handbook (US Geological Survey, 2016). Additive and multiplicative scaling factors, in addition to solar elevation angle, were provided by the image metadata. Scenes were processed using the "raster" package available for R software (Hijmans, 2016). ArcGIS software was then used to mosaic the images with a blending operator for overlapping pixels (ESRI, 2015).

Mosaicked images were finally processed, again using the "raster" package in R, to produce a modified soil vegetation index (MSAVI2) layer. This index was chosen for its increased sensitivity to vegetation in areas with exposed surface soil or rock fragments (Baugh and Groeneveld, 2006; Qi et al., 1994).

Both the river complexes and the slopes of the Brooks Range Mountains have a high percentage of exposed rock fragments that make this index appropriate. The determination of vegetated and non-vegetated pixels are critical in this project to distinguish between classified soil components and miscellaneous areas (e.g. rock outcrops or gravel bars). However, MSAVI2 is more sensitive to shadows in high relief areas when compared to the more commonly used normalized difference vegetation index (NDVI) (Zhang et al., 2015). As a result, we may expect lower accuracy on the shaded mountain slopes in the southern portion of the mapping study area.

### **2.3.3 Elevation Data**

The Digital Terrain Model (DTM) elevation data derived from the STAR-3 airborne Interferometric Synthetic Aperture Radar (IFSAR) was determined to be the highest resolution freely available data for this study area ([Elevation Portal], n.d.).

A total of 48 DTM tiles were merged together using the "raster" package in R (Conrad et al., 2015; Hijmans, 2016). The resulting merged DTM was then resampled from the original 5m to 10m, 30m, and 100m pixel size using a bilinear interpolation method. Finally, all DTMs were processed to fill surface depressions and to preserve downward slopes along flow paths for potential hydrologic modeling. This was done using the "Fill Sinks" module in SAGA software, with a 0.1% slope threshold (Wang and Liu, 2006). The resulting 10m, 30m, and 100m filled DTMs were used as snap rasters for all subsequent geoprocessing at each resolution, and all other raster layers were clipped to match their extent.

### **2.3.4 Geologic Data**

Since the study area spans multiple geologic units in the Brooks Range and Arctic Foothills, categorical geologic data were included in the modeling dataset. The geodatabase for the latest USGS generalized geologic map compilation was used for this project (Wilson et al., 2015). Map unit polygons were converted to raster format using ArcGIS "Polygon to Raster" conversion tool with outputs at 10m, 30m, and 100m pixel sizes (ESRI, 2015).

### **2.3.5 Geomorphons**

Landform classification systems derived from DTMs have been used extensively in geomorphological mapping, terrain modeling, and digital soil mapping (Bishop et al., 2012; McBratney et al., 2003; Mulder et al., 2011). For digital soil mapping, existing landform data can be incorporated into the project dataset as a categorical data layer (Cambule et al., 2013; Scull et al., 2005), transformed into a quantitative ruleset for modeling with continuous data layers (Nauman et



al., 2012; Nauman and Thompson, 2014) or new landform data can be developed specifically for the study area (Bacon et al., 2010; Hengl and Rossiter, 2003).

Recently, an unsupervised approach using "geomorphons" has been used with success in a wide variety of digital mapping and geomorphological research, including digital soil mapping (Ashtekar et al., 2014; Frankl et al., 2016; Jasiewicz et al., 2013; Libohova et al., 2016). Instead of fixed windows of analysis or neighborhood functions, this tool uses a flexible "viewshed" surrounding a given pixel. This approach is therefore pattern-based and scale-independent, which could be an advantage in this study area that encompasses a wide variety of landforms not resolvable at the same scales. For example, classifying summits, slopes, and valleys may be successful at one neighborhood size in the mountains, while these features may be best resolved using another neighborhood size in the rolling hills or plains. Inclusion of this layer was therefore based on a desire to include basic landform classification in the analysis without explicitly stratifying the landscape into discrete units based on relief and landform size.

A geomorphons layer was created using GRASS GIS software and the Geomorphons addon (Stepinski and Jasiewicz, 2011; Jasiewicz and Stepinski, 2013; GRASS Development Team, 2016). To ensure that broad landforms were adequately represented at the mapping scale of 1:63,360, a large maximum lookup value ( $L=1000$ ) was used on the 5m pixel size DTM before resampling to 10m, 30m and 100m for analysis (Jasiewicz and Stepinski, 2013). Smaller  $L$  values were tried and resulted in the over-definition of small features (especially in valley bottoms), which were inappropriate for the scale of mapping.

### **2.3.6 Terrain Attributes**

Digital soil mapping research has used a wide variety of terrain attributes derived from DTMs to serve as proxies for site characteristics and conceptual soil-forming factors. See McBratney et al., (2003) and Mulder et al., (2011) for reviews of modeling inputs used in contemporary digital soil mapping projects, and Bishop et al., (2012) for a review of terrain

attributes used in general geomorphological mapping.

The filled DTM described above was used to create a number of primary and secondary terrain attributes (Appendix B) using the "RSAGA" package in R (Brenning, 2008). This package uses modules from the SAGA software controlled via the R command line (Conrad et al., 2015). In total, 7 layers were chosen from a list of over 20 attributes. Inclusion was based on layer use in previously published digital soil mapping research (see reviews above), on the author's visual inspection of the layers viewed over a hillshade raster, and on previous characterization of soil forming factors in the arctic environment (Johnson et al., 2011; Ping et al., 2004; Ping et al., 2008b). Terrain ruggedness index (TRI) was calculated at multiple neighborhood sizes (10 cell and 25 cell radius) in an explicit attempt to stratify the study area into mountains, foothills, and plains without using categorical data.

### **2.3.7 Layer Stacking, Principal Component Analysis, and Band Statistics**

In the final step of data preprocessing, all raster layers were stacked for modeling (Table 2.2). Where necessary, extents were clipped and rasters were snapped to the filled DTMs at the appropriate resolution.

The three final stacks (at 10m, 30m, and 100m resolution) were then run through a principal component analysis (PCA) with corresponding band statistics outputs using ArcGIS. Only the 8 bands with non-categorical data were included in the analysis. The resulting correlation matrices from all resolutions showed high values (0.80 - 0.97) for slope and TRI layers only. This is to be expected when a first-order terrain attribute (such as slope) is used heavily in calculating a second-order attribute. All other layers were not highly correlated. For all resolutions, accumulative eigenvalues showed over 97% of the variance in the continuous dataset to be present in the first two PCA bands. To test whether inclusion of all terrain attributes is redundant, additional stacks replacing terrain attribute layers with the first two PCA bands were also used in modeling (stacks PCA-GM, PCA-GEO, and PCA-NOCAT).

## **2.4 Methods**

### **2.4.1 Random Forest Classification**

When compared to soil property predictive functions using geostatistical methods, such as kriging or cokriging (Heuvelink and Webster, 2001; McBratney et al., 2011; Webster and Oliver, 2001), there are generally fewer examples of soil classification functions. Reviews of recent digital soil mapping research show the percentage of papers focused on predictive mapping of soil classes ranged from 15.6% (Grunwald, 2009), 30% (McBratney et al., 2003), to "almost equally distributed" with soil property predictions (Lagacherie, 2008).

Functions used for predicting soil classes from environmental covariates include both linear models like logistic regression (Giasson et al., 2008; Hengl et al., 2007, 2014; Kempen et al., 2009), and complex machine learning functions like artificial neural networks (Behrens et al., 2005; Calderano Filho et al., 2014; Du et al., 2008; Moonjun et al., 2010; Zhu, 2000), and various decision trees (Connolly et al., 2007; Grinand et al., 2008; Pastick et al., 2014).

A random forest classification method was chosen here both for its use in previous soil and ecological mapping research (Chan and Paelinckx, 2008; Hengl et al., 2015; Rodriguez-Galiano et al., 2012; Roecker, et al., 2010; Stum et al., 2010), and for the option of using all available data for accuracy assessment (described below). The pedon dataset was also quite imbalanced with regard to classes (some classes were represented by less than 3 points), and the random forest method has been shown to perform well with imbalanced data when compared to other machine learning classifiers (Khoshgoftaar, 2007). The random forest machine learning algorithm takes the ensemble approach to decision tree modeling and includes at least two layers of randomness in each tree. Model training data (in this case, soil pedon classifications and their spatial locations) are bootstrapped for training each individual tree, while the nodes of each individual tree are also limited by using a small, random selection of layers within the input layer stack (Breiman, 2001; Liaw and Wiener, 2002). Final predicted classes are decided by receiving the majority of votes from all trees.

Processing was completed with the "raster" and "randomForest" packages for R, the latter built around the original Breiman et al. (1984) CART classification trees and Breiman (2001) random forest algorithm (Hijmans, 2016; Liaw and Wiener, 2002, 2008; R Core Team, 2016). A nested looping script was used to cycle through the multiple class columns in the original pedon data table, and through multiple layer stacks at each resolution (10m, 30m, and 100m).

Though the random forest method is not known for overfitting, a high percentage of irrelevant data in the input layer stack can cause poor performance of the model, especially when the selected number of variables to try at each node is small (Hastie et al., 2009). Therefore categorical data layers were varied in each stack in order to statistically compare (via error rate) if automated landform classification layers or geological layers were adding noise to the model, and also to qualitatively compare (via visual inspection) if the model outputs were skewed towards categorical data. Additional test stack combinations replaced the terrain attributes in the original stack with their first two principal components to test whether the inclusion of all terrain attributes was redundant and contributing noise to model inputs.

Finally, the random forest model runs were completed with 500, 750, and 1000 trees. Following Khoshgoftaar (2007), the formula " $\log_2 M + 1$ " was used for the median M value to determine the number of input layers tried at each node (mtry=3).

#### **2.4.2 Accuracy Assessment**

Because each individual tree is built using a random selection of ground-truth soil class data (a bootstrap sample), the random forest method eliminates the need for setting aside a percentage of our already limited data points for validation (Breiman, 2001). In contrast to the historical shortcomings of soil class mapping (Brevik et al., 2016), this modern classification method allows for an internal quantitative accuracy assessment by aggregating the "out-of-bag" (OOB) error rates from each individual tree (Liaw and Wiener, 2002). This estimate of OOB error is built in to the randomForest package and accompanies each modeled map output, and was used in this study to

evaluate model results. Since a variety of modeling resolutions and classification structures were tried, it was very simple to search the more than 1,000 output reports for the lowest error rates and compare modeling schema.

Maps were also evaluated visually by three independent soil scientists with soil mapping expertise in Alaska. Due to the large number of outputs, all maps could not be analyzed visually to compare results. To assure reasonable relationships between predicted soil classes and landforms, visual inspection of the best 5 results per taxonomic class was performed at 1:63,360 scales using the hillshade layer at a resolution identical to the map result. A mosaic of Landsat 8 imagery was also used as an overlay. A subsection of each physiographic region was presented to the evaluators. To avoid asking evaluators to judge specific soil classes in unfamiliar terrain (especially the Series Concept and Particle Size Class), instructions were to focus evaluations on whether landform and landform positions were adequately represented.

Maps were scored using a simple rubric divided according to landscape and landform assemblages common in each physiographic unit (Appendix A). Quantitative ratings of 1-5 were given in each category and the results totaled to compare between map results, with a maximum possible score of 60 for each mapset.

## **2.5 Results**

For each combination of taxonomic class and resolution, results were plotted by stack used and number of trees. No clear trend was observed between number of trees used to build the random forest and the resulting accuracy, but stack used and resolution did appear to affect accuracies (Figure 2.4). Results are briefly summarized by taxonomic class below.

### 2.5.1 Estimated OOB Accuracy by Taxonomic Class and Resolution

When predicting soil order, at all resolutions stacks using only categorical data ("catonly") performed worse than all other stacks. Accuracies ~65-75% were observed at all resolutions for all stacks. PCA stacks measured slightly lower than non-PCA stacks at 100m resolution but still had accuracies above 65%. PCA stacks performed best at 30m resolution. Overall, non-PCA stacks performed only slightly better than PCA stacks. When predicting soil suborder, at all resolutions stacks using only categorical data ("catonly") performed worse than all other stacks. Accuracies ~55-65% were observed at all 10m and 30m resolutions for non-PCA stacks, with PCA stacks slightly lower at ~50-55%. At 100m resolution, all stacks (besides "catonly") performed equally with accuracies ~50-60%.

When predicting soil great group, at all resolutions stacks using only categorical data ("catonly") performed worse than all other stacks. Accuracies ~45-50% were observed at all 10m and 30m resolutions for non-PCA stacks, with PCA stacks slightly lower at ~35-45%. At 100m resolution, accuracies ~40-45% were observed for non-PCA stacks, with PCA stacks slightly lower at ~35-40%.

When predicting series concept, at all resolutions stacks using only categorical data ("catonly") performed worse than all other stacks. Accuracies ~45-50% were observed in non-PCA stacks at 30m resolution, with PCA stacks slightly lower at ~35-40%. At 10m resolution, accuracies were ~40-45% for non-PCA stacks, with PCA stacks slightly lower ~30-40%. At 100m resolution, accuracies were observed ~35-40% for non-PCA bands, and a slightly lower (but wider range of) accuracies ~30-40% for PCA bands.

When predicting particle size class, at all resolutions stacks using only categorical data ("catonly") performed worse than all other stacks. Accuracies ~55-65% were observed in non-PCA stacks at 10m and 30m resolutions, with PCA stacks ~45-50%. At 100m resolution, PCA stacks performed equal to or slightly better than non-PCA stacks, with accuracies of ~50-60% and ~45-55%, respectively. This is the only taxonomic class for which PCA stacks performed equal to or

better than non-PCA stacks. When plotted by number of classes in each taxonomic group, the data suggest a trend of decreasing accuracy and kappa values with increasing number of predicted classes (Figure 2.4a). There is no obvious trend when plotting accuracy and kappa against grid size or number of trees used in modeling (Figure 2.4b and 2.4c).

### **2.5.2 Visual Inspection of Map Results**

Overall, map ratings varied considerably between evaluators. Two evaluators had an almost equal range of total map scores (21 to 48 points out of 60, and 21 to 47 points out of 60, respectively) and a third evaluator had consistently higher scores (35 to 60 points out of 60). Surprisingly, evaluators seemed to prefer the maps with lowest estimated OOB accuracy when total map scores were averaged for each map set (Table 2.3). Top rated maps predicted soil Series Concept or soil Great Group, and were all assessed to be less than 50% accurate. The maps with lowest ratings predicted either soil Order or Great Group, but were almost all above 70% accuracy.

The three physiographic regions were also evaluated individually. Only the mountainous subsection was rated highly by evaluators when estimated OOB accuracy was low. The foothills and coastal plain subsections were more variable when comparing evaluator ratings to accuracy values, with no obvious trend.

## **2.6 Discussion**

### **2.6.1 Stacks**

Replacing terrain attributes with first and second PCA bands lowered accuracy in all stack/resolution combinations but one. This was not surprising, as the PCA correlation matrices for all resolutions showed fairly low values between all input bands except for slope and TRI index. Reducing the dimensionality of the dataset with PCA lowers accuracy ~5-10% without a significant decrease in processing time for a dataset of this size. For this reason, PCA analysis can be

considered an unnecessary step with no added value in this particular pilot study. However, it should be noted that the "PCA-GEO" layer (which includes the categorical geology layer) performed best in most cases.

The issue of finding a classification method that allowed the use of categorical landform and geologic data was important in the decision to try random forest modeling in this study. However, results show that using no categorical data at all did not result in lower accuracies, and using only categorical data in the classification provides much lower accuracies. This is surprising as most soil mapping approaches recognize soil-landform and soil-parent material relationships as major factors in developing soil components and map units. Either these categorical layers are poor representations of their respective phenomena (and therefore the random forest algorithm did not make use of them in the classification process), or continuous terrain attributes and vegetation indices are simply more powerful predictors of soil classes.

In map results with the highest accuracy by class, variable importance measures computed in the randomForest package show less than 2% mean decrease in accuracy across all trees when categorical variables are permuted (Archer and Kimes, 2008). From this we can infer that the most successful classification trees depend mostly on continuous data in all stacks. Like PCA, automated landform analysis is an extra step with little added value in this study. Removing landform and geology layers from the analysis would eliminate the use of extra software to create individual layers (e.g. using GRASS for geomorphons) and allow for direct comparison of random forest modeling with other classification methods that do not allow for categorical data inputs.

### **2.6.2 Resolution**

With regards to modeling resolution and accuracy, 10m and 30m show comparable results for the vast majority of cases with 100m performing ~5-10% lower. While it may seem intuitive that higher resolution datasets performed better, interpretation of these results is confounded by the fact that the study area includes both high and low relief landscapes and accuracy is computed for the



study area as a whole. Thompson et al., (2001) and Pain et al., (2005) have shown that high resolution DEMs may have the greatest positive effect on map accuracy when used in high relief areas, while Cavazzi et al., (2013) actually noted a decrease in random forest model accuracy using high resolution DEM in low relief areas.

In addition to model resolution, Maynard and Johnson (2014) found window size to have a larger impact on soil property predictions than DEM resolution when using a high resolution (1-5m) dataset. Roecker and Thompson (2010) determined that terrain attributes that compute surface curvatures are most sensitive to window size, with optimal windows determined with reference to the size of local landforms present in the mapping area. Window sizes were not investigated systematically in this study, though it is a promising direction for future research. The inclusion of local mean filters alongside raw covariate data by Moran and Bui (2002) also provides an example of using neighborhood analysis instead of varying DEM resolution. The challenge in classifying a large area that varies from high to low relief is to find resolutions and window sizes that provide adequate detail where needed without overanalyzing the terrain in other areas.

The idea of physiographic stratification was not overlooked in this study; the intent was to test if the random forest method could perform a meaningful landscape stratification from the data provided. Our visual inspection (detailed below) shows that soil classes generally appear where they should be geomorphically. Most geomorphic misclassification is observed where soils exclusively located on mountain slopes are predicted in lower hillslope areas, or where soils exclusively located on floodplains appear in upland areas. In the future, incorporating local mean filters or outright geomorphic stratification might improve accuracy while using moderate resolution DTMs.

### **2.6.3 Classification Schema**

Accuracies of model results were closely related to the level of taxonomy used to classify the input point data. This is likely due to increasingly narrow levels of taxonomy having more classes. For example, there are 3 classes at the Order level, 8 at the Suborder level, and 15 at the

Great Group level (excluding non-soil classes of water and gravel). As the number of classes to predict increases, accuracy appears to decrease. The question of which classification scheme is best will no doubt depend on the final use of the map. As a general guideline, map producers using similar methods should expect accuracies of less than 50% when attempting to predict 15 or more classes.

#### **2.6.4 Trees**

There is no clear pattern when plotting model accuracy against number of trees. Using a laptop PC with Intel Core i7-6700HQ Processor at 2.6GHz and 16GB of RAM, the difference between processing 100 and 1000 trees is over 5 hours. In this case, processing time is probably the deciding factor when considering an optimal number of trees for a classification.

#### **2.6.5 Pedon Data**

It is unclear how the model accuracy was affected by point data distribution or quantity. Since this study used a mixture of legacy pedon data and landform-based transect data, input point data was both unevenly distributed in geographical space and of very low density. Random forest algorithms rely on spatial association of soil properties or classes to environmental covariates throughout the feature space, with each point evaluated individually from adjacent points. As such, random forests are not necessarily sensitive to the geographic distribution of point data, but are sensitive to point data distribution within the feature space. As detailed in Bui et al., (2006), prediction of soil classes in unobserved areas can fail when sampling design does not include representative areas for the entire feature space.

In contrast, spatial interpolation methods (kriging or cokriging) rely on the spatial autocorrelation of observations and therefore perform best with high-density sampling in geographic space. Predictions are also typically not extended beyond the sampling extent. Miller et

al., (2016) found that multiple logistic regression performs better than kriging or cokriging when modeling soil properties outside of the sampling extent, suggesting that even among geostatistical methods spatial autocorrelation may be a poor choice when mapping remote areas.

Brungard and Boettinger (2010) used a conditioned Latin Hypercube Sampling (cLHS) method to determine an optimal sampling size of 200-300 points to adequately represent the feature space in a ~300km<sup>2</sup> study area. While the approach is admirable, this level of sampling density (1 to 1.5 observations per km<sup>2</sup>) is logistically implausible in almost all unmapped areas of Alaska. Liess (2015) raises the issue of sampling design in digital soil mapping DSM where using proper geostatistical sampling methods is not possible, and suggests an optimization process that starts with determining the accessible zones within the mapping area and dividing the feature space within those accessible zones. In a similar fashion, Roudier et al., (2012) incorporate a “cost” layer into the cLHS optimization. A hybrid approach including legacy data alongside an optimized sampling design would be required for this specific project where some data exists but may not be entirely representative.

It is likely that point data distribution may affect model results more than point data quantity. For example, when using a boosted classification tree, Grinand (2008) did not find an increase in classification accuracy of external areas by increasing internal sampling density. More methods to optimize sampling design in a machine learning context should be explored in the future, with special attention paid to remote areas with difficult access.

The difficulty of traveling across tundra on foot also contributed to clustering of point data locations, with soil scientists choosing to transect regions where a maximum of landform variability could be observed with a minimum of effort. This method is efficient in the field, but often causes two unique soil classes to be observed in adjacent pixels, or even within one pixel when using coarse resolution data. This poor spacing of observations and possible inappropriate description methods highlight the need for an accepted soil sampling protocol for patterned ground features, like those proposed by Ping et al., (2013).

Beyond training the model, point data is also used to evaluate model accuracies. When predicting soil properties, Bishop et al., (2015) show that using point-based ground truth data for validation often presents a worst-case scenario for map accuracy. Mean values derived from block supports are shown to have higher accuracy ratings, with supports based on grids or polygons surrounding point-based data. This method has not yet been incorporated into the random forest accuracy assessment and would likely require a separate validation dataset, which is limiting in the context of mapping remote areas with logistical constraints and small input datasets.

Point-based accuracy assessments have also been used to compare DSM products with conventional soil survey maps. Zhu et al. (2001) show that a SoLIM-derived DSM product is able to predict both soil classes and properties at a higher accuracy than a conventional polygon-based soil survey map, though the fairness of comparing raster and vector products using individual points is debatable for a few reasons.

The raster map is able to show variation over much shorter distances and also can include small "islands" of contrasting values within larger homogenous zones in the raster. These would be treated as inclusions in soil polygons, and may cause the map to appear less accurate when using individual points for evaluation. This is especially likely if only major components of the mapunit are considered.

Conventional soil map units also generally focus on landforms and landform positions and were not designed to be an accurate representation of soils at any given point on the map, but instead were tailored for specific use and management applications (MacMillan, 2008). To assess conventional maps using point supports is to assess the product on a task it was never intended to accomplish.

In response to these shortcomings in accuracy assessment, MacMillan (2008) reviews several neighborhood functions that compute the proportion of correct classes within a given number of cells. These methods may be more appropriate when comparing competing DSM products or comparing DSM products with rasterized conventional maps. Fuzzy matching

techniques and other methods not dependent on exact class matches in point-based accuracy assessments would be ideal for internal evaluation of DSM products from the use and management perspective, especially in cases where multiple soil classes may have similar interpretive value to the map user. Thresholds for predictive accuracy would need to be developed by project leaders and would depend on the intended use of the soil survey.

More research is needed to determine the proper validation method for regional-scale soil maps and come to a consensus on acceptable accuracy levels for various uses of soil maps. As discussed by Baveye and Laba (2015), we must also ask what degree of heterogeneity needs to be conveyed to the end user, and at what confidence, when considering the interpretive value of soil class or soil property maps.

#### **2.6.6 Map Quality**

Notably, the maps at 100m resolution performed worst in visual evaluations. As the largest pixel size in the mapset, these maps may have suffered when directly compared to higher resolution maps viewed during evaluation. These maps also had fewer classes (only 6 at the taxonomic level of Order) than the best performing Series Concept maps. This may suggest a bias towards maps with more unique classes, as individual landform positions may appear more explicit. The finer-scale map with more classes may simply be more visually appealing as it appears to have more information. Soil scientists do not often create or use maps depicting taxonomic levels as coarse as soil Order, and the experience of viewing a taxonomically coarse classification scheme at 1:63,360 scale may have contributed to low visual evaluation ratings.

Though maps predicting Series Concept performed worst in accuracy values derived from point data, their high performance in visual evaluation shows that landforms and landform positions have been modeled realistically. Overall, plausible spatial distribution of soil classes may be more important than class accuracy, considering that any soil class consistently associated with a landform in the map result could be correlated to a more appropriate class (or classes) during post-

processing if necessary.

The apparent disagreement between model accuracy and visual ratings suggests that OOB estimates of accuracy should not be the sole method of evaluating DSM products. Visual evaluation has been used extensively in traditional soil mapping as part of standard quality control workflows, but it is hardly mentioned in the DSM literature. In a recent review of DSM approaches, Grunwald (2009) did not even list visual inspection as a method of accuracy assessment, possibly because of its subjective nature.

#### **2.6.6.1 Expert Knowledge Validation**

Expert or tacit knowledge is most often considered during the model building stage and is incorporated via the selection of specific predictive variables, pruning of pedon data from the training data set (i.e., choosing a modal dataset), and additions of an expert's remotely sensed observations to train the model (Kienast-Brown et al., 2017). Expert knowledge is also considered as an abstract body of undocumented rules and assumptions that exist in legacy soil survey data and can potentially be translated into "knowledge-based" or "rule-based" quantitative models (Jensen, 2005; Zhu et al., 2001). DSM using expert knowledge-based classification has been shown to be successful (McKay et al., 2010; Nauman and Thompson, 2014; Shi et al., 2004), but expert knowledge-based accuracy assessment is rarely reported.

Using expert knowledge for model validation moves in the opposite direction from quantitative to qualitative assessment, and is essentially unrepresented in published DSM workflows. Recent soil surveys using DSM have used expert knowledge to rate the results of different modeling approaches (Cole and Boettinger, 2006; McKay et al., 2010; Shi et al., 2009; T. D'Avello, personal communication, 2018).

In the DSM soil survey of Essex County, Vermont, soil map units already delineated by soil scientists were compared to the DSM product in order to evaluate how a rule-based model

compared to subjective spatial predictions (McKay et al., 2010; Shi et al., 2009). This method may be appropriate for small survey areas with few components, but would be difficult to employ in large survey areas with multiple geomorphic environments. The expert's manual delineations would ideally need to include most mapunits or components in order to truly test the model results. For this reason, a stratification process for selecting test areas within the soil survey would need to be developed to minimize the time spent manually digitizing while maximizing map unit or component variability.

In the DSM soil survey of the Boundary Waters Canoe Area, Minnesota, local experts reviewed a variety of model results from different methods (random forest, unsupervised classification, rule-based classification, etc.) and selected the best results by class (T. D'Avello, personal communication, 2018). These class results were then combined into a hybrid raster map that was evaluated quantitatively. In this methodology, expert knowledge was used as a map validation and model building resource in the same step. For all the quantitative DSM methods employed during that project, final map results were therefore heavily influenced by expert opinion with selection criteria remaining implicit.

These workflows risk negating two of the key perceived benefits of DSM - a) the elimination of time-consuming heads-up digitizing and subjective spatial prediction of soil types, and b) the minimization of implicit biases and tacit knowledge in soil mapping. However, when viewed against the lack of expert knowledge validation workflows in published DSM literature, these studies provide useful examples of incorporating expert knowledge at the assessment stage. In the future, validation criteria should ideally be made more explicit to allow an attempt at reproducibility and reflect the goals of DSM to provide rigorous, quantitative, reproducible models.

As the criteria for expert-knowledge validation varies significantly between soil survey areas, it is unlikely that a standardized set of criteria or a common validation workflow will be proposed at this time. But the results of this research indicate that the current focus on quantitative accuracy assessments alone is not sufficient to produce quality soil maps. The degree to which

expert knowledge should be employed in map validation and the criteria by which maps should be evaluated is therefore open for debate in future DSM research. The visual evaluation rubric used in this research is one option for providing form and structure to what is by definition a subjective process (Appendix A).

When considering DSM in Alaska, it is important to note that any discussion of expert knowledge validation must recognize that many of the remote areas that will be represented by DSM products have not been previously characterized by soil scientists. Though local experts do have some predictive ability based on patterns observed elsewhere, the state is essentially *terra incognita* when considering fine to medium scale soil component data. This challenge cannot be ignored as the soil scientists working in the field for only a few months may have the most experience in a given area, and will be considered experts in evaluating DSM products covering millions of acres. It is interesting to note in this study that the map results were consistently rated highest by the soil scientist with the most field experience in the mapping area. This speaks to the nature of soil survey as a process where tacit knowledge and local field experience are crucial in both the creation and evaluation of soil maps.

There is considerable opportunity to use expert knowledge combined with other remote sensing strategies for rigorous accuracy assessments. However, we can expect this process to be difficult if digital soil maps are presented in raster form and/or on a component scale (as in this research), since most soil scientists are familiar with vector-based aggregated map products. As stated above, point-based accuracy assessments may not be the best choice for DSM, and the author recommends incorporating more expert knowledge and other alternative methods of model assessment into future DSM research.



#### **2.6.6.2 Perception Bias and Relationship to Scale**

Most soil mappers are trained and comfortable in associating soil types and properties with landforms and slope positions within a survey area. However, the evaluation of soil maps based on more generalized taxonomic criteria, that are not necessarily related to specific criteria used for mapping soil types and properties, poses a challenge to the soil surveyor as shown by the results of the expert knowledge based validation. The underlying confounding factor could be the scale discrepancies and the increased level of generalization from field based soil type evaluation to taxonomic based soil type evaluation. The framing of the questions for the expert knowledge validation could also be a factor. In this study, the goal of the expert knowledge based validation was to evaluate whether the mapped soil types conformed to landform position and shape when viewed over a hillshade and Landsat imagery and not whether or not the individual soil taxonomic units on the map were correct. The evaluation was targeted towards soil-landform relationships than taxonomy. Thus, the weight of soil scientist's subjective evaluation on the final DSM product should therefore be considered carefully, as project leaders will not have the luxury of consulting experts who have spent their entire careers working around a given soil survey area.

#### **2.6.7 Recommendations for Post-Processing**

Though it is not traditionally a taxonomic class in and of itself, particle size class (PSC) was predicted with relatively high accuracy, having only 9 unique classes. Combining predicted PSC with another map output via a raster calculator function may provide additional interpretive value without altering the classification methodology. Individual soil Subgroup prefixes were also predicted with a very high accuracy, though the group had only 7 unique classes and ~80% of the input data was in the class of "Typic". The resulting map of almost entirely one class was not very useful, and those results are not included above for that reason. However, one could theoretically predict the PSC, Subgroup prefix, and Great Group separately and combine the outputs into a higher taxonomic class (e.g. taxon above family). It would be difficult to determine the accuracy of

such a product, and some class combinations may be impossible taxonomic units. The result would likely appear closer to a traditional SSURGO level component legend and would potentially increase map utility or aid in comparisons against traditional soil survey products.

Obvious mapping errors could be reduced by reclassifying pixels within basic physiographic stratifications (for instance, pixels within the Arctic Coastal Plain that are predicted as components sampled exclusively in the Brooks Range Mountains). This might be appropriate for the most egregious errors, but if landscape stratification is to be performed at all one might suggest modeling each strata separately from the beginning rather than in post-processing. A more appropriate workflow might be to reclassify pixels as complexes of multiple components using conditional reasoning. When classes are often observed or predicted closely together in geographic space, the pixels could be reclassified as a complex of the two classes to reduce the "confetti" effect of modeling two soils that share similar landforms and landform positions (e.g. Historthels and Histoturbels in the Arctic Coastal Plain). Reclassification could be done when pixels of two specified classes are adjacent, or when pixels are within some distance of each other.

In land cover mapping, "noisy" model outputs are commonly processed through boundary cleaning, majority filtering, or other workflows involving expert knowledge and/or ancillary reference data (Rozenstein and Karnieli, 2011; Van de Voorde et al., 2007). When compared to continuous value rasters, classified outputs have fewer tools available as the processing is often conditional rather than arithmetic or statistical. Replacing single pixels or small clusters of pixels with neighboring classes can greatly enhance the visual appeal of the map and is a crucial step before converting the raster model to SSURGO-style polygons.

Finally, it is important to note the accuracy assessment performed by the random forest algorithm is valid for the raw model only. Ideally, a separate accuracy assessment would be performed for the finished, post-processed model. Combining classes into complexes or otherwise altering the original class structure would require a more complicated investigation of which validation points are misclassified. Again, workflows that rely on conditional statements may be most appropriate. When comparing to raw random forest accuracies, a k-fold cross-validation method may be the most simple and straightforward assessment of post-processed model accuracy in this case.

## **2.7 Conclusions**

Overall, this research suggests that random forest modeling is an appropriate method for digital soil mapping in the sparsely sampled regions of Alaska. At all resolutions tested, using a small number of trees (100 or less) on a simple stack of continuous environmental covariates provides accurate soil maps at taxonomic levels of great group and higher. The addition of categorical data did not substantially increase map accuracy and should be considered unnecessary, especially when inclusion of these layers requires additional software. Results shown here are a marked improvement from the currently available STATSGO2 dataset due to the finer scale model output and the increased number of data points used to populate the model.

Global climate modelers could benefit from a wider application of random forest digital soil mapping throughout the circumpolar North using coarse resolution outputs and a small number of classes representing basic soil taxonomic groups. However, more research is needed to determine the most reasonable number of classes required to provide adequate interpretive values for natural resource management on public and private lands. As with all soil surveys, stakeholders will need to discuss the scope of the project and determine the level of detail that will meet their needs.

Optimization workflows for future soil sampling in remote regions should be pursued, and should ideally include legacy data wherever possible. Future soil sampling should be carried out

with modeling resolution in mind to avoid sampling clusters, with additional efforts to consistently describe patterned ground and other periglacial features to improve model training.

## **2.8 Acknowledgments**

This research was supported by the USDA Natural Resources Conservation Service grant (Award # 68-7482-15-531) and the UADA NIFA program. We would like to thank colleagues Stephanie Schmit, Eric Geisler, and Matt Ferderbar who provided soils expertise that greatly assisted the research. We also thank Dr. Zamir Libohova (USDA-NRCS) and Dr. Jordi Cristobal (University of Alaska, Fairbanks) for comments that greatly improved the manuscript.

## **2.9 References**

- Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4), 2249–2260.  
<https://doi.org/10.1016/j.csda.2007.08.015>
- Ashtekar, J., Owens, P., Brown, R., Winzeler, H., Dorantes, M., Libohova, Z., ... Castro, A. (2014). Digital mapping of soil properties and associated uncertainties in the Llanos Orientales, South America. In D. Arrouays, N. McKenzie, J. Hempel, A. de Forges, & A. McBratney (Eds.), *GlobalSoilMap* (pp. 367–372). CRC Press. <https://doi.org/10.1201/b16500-67>
- Bacon, S. N., McDonald, E. V., Dalldorf, G. K., Baker, S. E., Sabol, D. E., Minor, T. B., ... Bullard, T. F. (2010). Predictive Soil Maps Based on Geomorphic Mapping, Remote Sensing, and Soil Databases in the Desert Southwest. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 411–421). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_32](https://doi.org/10.1007/978-90-481-8863-5_32)
- Baugh, W. M., & Groeneveld, D. P. (2006). Broadband vegetation index performance evaluated for a low-cover environment. *International Journal of Remote Sensing*, 27(21), 4715–4730.  
<https://doi.org/10.1080/01431160600758543>

- Baveye, P. C., & Laba, M. (2015). Moving away from the geostatistical lamppost: Why, where, and how does the spatial heterogeneity of soils matter? *Ecological Modelling*, 298, 24–38.  
<https://doi.org/10.1016/j.ecolmodel.2014.03.018>
- Behrens, T., Förster, H., Scholten, T., Steinrücken, U., Spies, E.-D., & Goldschmitt, M. (2005). Digital soil mapping using artificial neural networks. *Journal of Plant Nutrition and Soil Science*, 168(1), 21–33. <https://doi.org/10.1002/jpln.200421414>
- Bishop, M. P., James, L. A., Shroder, J. F., & Walsh, S. J. (2012). Geospatial technologies and digital geomorphological mapping: Concepts, issues and research. *Geomorphology*, 137(1), 5–26. <https://doi.org/10.1016/j.geomorph.2011.06.027>
- Bishop, T. F. A., Horta, A., & Karunaratne, S. B. (2015). Validation of digital soil maps at different spatial supports. *Geoderma*, 241–242, 238–249.  
<https://doi.org/10.1016/j.geoderma.2014.11.026>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.  
<https://doi.org/10.1023/A:1010933404324>
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees*. New York: Chapman and Hall.
- Brenning, A. (2008). Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. *SAGA—Seconds out*, 19, 23–32.
- Brevik, E. C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., & Jordán, A. (2016). Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma*, 264, 256–274. <https://doi.org/10.1016/j.geoderma.2015.05.017>
- Brungard, C. W., & Boettinger, J. L. (2010). Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 67–75). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_6](https://doi.org/10.1007/978-90-481-8863-5_6)

- Bui, E. N., Simon, D., Schoknecht, N., & Payne, A. (2006). Chapter 15 Adequate Prior Sampling is Everything: Lessons from the Ord River Basin, Australia. In *Developments in Soil Science* (Vol. 31, pp. 193–608). Elsevier. [https://doi.org/10.1016/S0166-2481\(06\)31015-X](https://doi.org/10.1016/S0166-2481(06)31015-X)
- Calderano Filho, B., Polivanov, H., Chagas, C. da S., Carvalho Júnior, W. de, Barroso, E. V., Guerra, A. J. T., & Calderano, S. B. (2014). Artificial neural networks applied for soil class prediction in mountainous landscape of the Serra do Mar1. *Revista Brasileira de Ciência Do Solo*, 38(6), 1681–1693. <https://doi.org/10.1590/S0100-06832014000600003>
- Cambule, A. H., Rossiter, D. G., & Stoorvogel, J. J. (2013). A methodology for digital soil mapping in poorly-accessible areas. *Geoderma*, 192, 341–353. <https://doi.org/10.1016/j.geoderma.2012.08.020>
- Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., & Fealy, R. (2013). Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma*, 195–196, 111–121. <https://doi.org/10.1016/j.geoderma.2012.11.020>
- Chan, J. C.-W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6), 2999–3011. <https://doi.org/10.1016/j.rse.2008.02.011>
- Cole, N. J., & Boettinger, J. L. (2006). Chapter 27 Pedogenic Understanding Raster Classification Methodology for Mapping Soils, Powder River Basin, Wyoming, USA. In A. B. M. and M. V. P. Lagacherie (Ed.), *Developments in Soil Science* (Vol. 31, pp. 377–619). Elsevier. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0166248106310276>
- Connolly, J., Holden, N. M., & Ward, S. M. (2007). Mapping Peatlands in Ireland using a Rule-Based Methodology and Digital Data. *Soil Science Society of America Journal*, 71(2), 492. <https://doi.org/10.2136/sssaj2006.0033>

- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J. (2015): System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, *Geosci. Model Dev.*, 8, 1991-2007.
- Deluca, T. H., & Boisvenue, C. (2012). Boreal forest soil carbon: distribution, function and modelling. *Forestry*, 85(2), 161–184. <https://doi.org/10.1093/forestry/cps003>
- Du, C., Linker, R., & Shaviv, A. (2008). Identification of agricultural Mediterranean soils using mid-infrared photoacoustic spectroscopy. *Geoderma*, 143(1–2), 85–90. <https://doi.org/10.1016/j.geoderma.2007.10.012>
- Elevation Portal. (n.d.). State of Alaska Division of Geological and Geophysical Surveys Website. Retrieved November, 2016. <https://elevation.alaska.gov>
- ESRI (Environmental Systems Research Institute), (2015). ArcGIS Release 10.4. Redlands, CA.
- Frankl, A., Lenaerts, T., Radusinoviæ, S., Spalevic, V., & Nyssen, J. (2016). The regional geomorphology of Montenegro mapped using Land Surface Parameters. *Zeitschrift Für Geomorphologie*, 60(1), 21–34. <https://doi.org/10.1127/zfg/2016/0221>
- Gallant, A. L., Binnian, E. F., Omernik, J. M., & Shasby, M. B. (1995). Ecoregions of Alaska (USGS Numbered Series No. 1567) (p. 78). U.S. Geological Survey. Retrieved from <http://pubs.er.usgs.gov/publication/pp1567>
- Giasson, E., Figueiredo, S. R., Tornquist, C. G., & Clarke, R. T. (2008). Digital soil mapping using logistic regression on terrain parameters for several ecological regions in Southern Brazil. In *Digital soil mapping with limited data* (pp. 225-232). Springer, Dordrecht.
- GRASS Development Team. (2016). Geographic Resources Analysis Support System (GRASS) Software, Version 6.4. Open Source Geospatial Foundation. <http://grass.osgeo.org>.

- Grinand, C., Arrouays, D., Laroche, B., & Martin, M. P. (2008). Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, 143(1–2), 180–190.  
<https://doi.org/10.1016/j.geoderma.2007.11.004>
- Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, 152(3–4), 195–207. <https://doi.org/10.1016/j.geoderma.2009.06.003>
- Hamilton, T. D. (2003). Glacial geology of the Toolik Lake and upper Kuparuk River regions. Institute of Arctic Biology. University of Alaska Fairbanks.
- Hastie, Trevor ; Tibshirani, Robert & Friedman, Jerome (2009). *The Elements of Statistical Learning*. Springer: New York.
- Hengl, T., & Rossiter, D. G. (2003). Supervised landform classification to enhance and replace photo-interpretation in semi-detailed soil survey. *Soil Science Society of America Journal*, 67(6), 1810–1822.
- Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., ... Gonzalez, M. R. (2014). SoilGrids1km — Global Soil Information Based on Automated Mapping. *PLoS ONE*, 9(8), e105992. <https://doi.org/10.1371/journal.pone.0105992>
- Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., ... Tondoh, J. E. (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *PLOS ONE*, 10(6), e0125814.  
<https://doi.org/10.1371/journal.pone.0125814>
- Hengl, T., Toomanian, N., Reuter, H. I., & Malakouti, M. J. (2007). Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. *Geoderma*, 140(4), 417–427. <https://doi.org/10.1016/j.geoderma.2007.04.022>
- Heuvelink, G. B. ., & Webster, R. (2001). Modelling soil variation: past, present, and future. *Geoderma*, 100(3–4), 269–301. [https://doi.org/10.1016/S0016-7061\(01\)00025-8](https://doi.org/10.1016/S0016-7061(01)00025-8)



- Hijmans, R. (2016). raster: Geographic Data Analysis and Modeling. <https://CRAN.R-project.org/package=raster>
- Hinkel, K. M., Eisner, W. R., Bockheim, J. G., Nelson, F. E., Peterson, K. M., & Dai, X. (2003). Spatial Extent, Age, and Carbon Stocks in Drained Thaw Lake Basins on the Barrow Peninsula, Alaska. *Arctic, Antarctic, and Alpine Research*, 35(3), 291–300. [https://doi.org/10.1657/1523-0430\(2003\)035\[0291:SEAACS\]2.0.CO;2](https://doi.org/10.1657/1523-0430(2003)035[0291:SEAACS]2.0.CO;2)
- Hugelius, G. (2012). Northern Circumpolar Soil Carbon Database. ECDS Environment Climate Data Sweden. <https://doi.org/10.5879/ecds/000000001>
- Hugelius, G., Strauss, J., Zubrzycki, S., Harden, J. W., Schuur, E. A. G., Ping, C.-L., ... Kuhry, P. (2014). Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps. *Biogeosciences*, 11(23), 6573–6593. <https://doi.org/10.5194/bg-11-6573-2014>
- Huryn, A. D., & Hobbie, J. E. (2012). Land of extremes: a natural history of the Arctic North Slope of Alaska. Fairbanks, Alaska: University of Alaska Press.
- Jafarov, E. E., Marchenko, S. S., & Romanovsky, V. E. (2012). Numerical modeling of permafrost dynamics in Alaska using a high spatial resolution dataset. *The Cryosphere*, 6(3), 613–624. <https://doi.org/10.5194/tc-6-613-2012>
- Jasiewicz, J., & Stepinski, T. F. (2013). Geomorphons — a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182, 147–156. <https://doi.org/10.1016/j.geomorph.2012.11.005>
- Jasiewicz, J., Netzel, P., & Stepinski, T. F. (2013). Content-based landscape retrieval using geomorphons. *Geomorphometry* 2013.
- Jensen, J.R. (2005). Introductory digital image processing: A remote sensing perspective, 3rd edition. Pearson Prentice Hall, pp. 296-300, 301-321, 315-316.

- Johnson, K. D., Harden, J., McGuire, A. D., Bliss, N. B., Bockheim, J. G., Clark, M., ... Valentine, D. W. (2011). Soil carbon distribution in Alaska in relation to soil-forming factors. *Geoderma*, 167–168, 71–84. <https://doi.org/10.1016/j.geoderma.2011.10.006>
- Jorgenson, M. T., Roth, J. E., Emers, M., Schlentner, S. F., Swanson, D. K., Pullman, E. R., & Stickney, A. A. (2003). An ecological land survey in the Northeast Planning Area of the National Petroleum Reserve–Alaska, 2002. ABR. Inc., Fairbanks, AK.
- Jorgenson, M. T., Roth, J. E., Miller, P. F., Macander, M., Duffy, M. S., Wells, A., ... Pullman, E. R. (2009). An ecological land survey and landcover map of the Arctic Network (Natural Resource Technical Report No. NPS/ARC/NRTR—2009/270). Fort Collins, Colorado: National Park Service. Retrieved from [http://science.nature.nps.gov/im/units/arcn/documents/documents/NPS\\_ARCN\\_NRTR-2009-270\\_EcologicalLandSurveyLandcoverMap.pdf](http://science.nature.nps.gov/im/units/arcn/documents/documents/NPS_ARCN_NRTR-2009-270_EcologicalLandSurveyLandcoverMap.pdf)
- Kempen, B., Brus, D. J., Heuvelink, G. B. M., & Stoorvogel, J. J. (2009). Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma*, 151(3–4), 311–326. <https://doi.org/10.1016/j.geoderma.2009.04.023>
- Khoshgoftaar, T. M., Golawala, M., & Hulse, J. V. (2007). An Empirical Study of Learning from Imbalanced Data Using Random Forest (pp. 310–317). IEEE. <https://doi.org/10.1109/ICTAI.2007.46>
- Kienast-Brown, S., Libohova, Z., & Boettinger, J. (2017). Chapter 5: Digital Soil Mapping. In *Soil survey manual*. 2017. C. Ditzler, K. Scheffe, and H.C. Monger (eds.). USDA Handbook 18. Government Printing Office, Washington, D.C.
- Lagacherie, P. (2008). Digital Soil Mapping: A State of the Art. In A. E. Hartemink, A. McBratney, & M. de L. Mendonça-Santos (Eds.), *Digital Soil Mapping with Limited Data* (pp. 3–14). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-1-4020-8592-5\\_1](https://doi.org/10.1007/978-1-4020-8592-5_1).
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18–22.

- Liaw, A., & Wiener, M. (2008). randomForest: Breiman and Cutler's Random Forests for Classification and Regression. <http://CRAN.R-project.org/package=randomForest>
- Libohova, Z., Winzeler, H. E., Lee, B., Schoeneberger, P. J., Datta, J., & Owens, P. R. (2016). Geomorphons: Landform and property predictions in a glacial moraine in Indiana landscapes. *CATENA*, 142, 66–76. <https://doi.org/10.1016/j.catena.2016.01.002>
- Liess, M. (2015). Sampling for regression-based digital soil mapping: Closing the gap between statistical desires and operational applicability. *Spatial Statistics*, 13, 106–122. <https://doi.org/10.1016/j.spasta.2015.06.002>
- MacMillan, R. A. (2008). Experiences with applied DSM: protocol, availability, quality and capacity building. In *Digital soil mapping with limited data* (pp. 113-135). Springer, Dordrecht.
- Maynard, J. J., & Johnson, M. G. (2014). Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: Effects of grid resolution vs. neighborhood extent. *Geoderma*, 230–231, 29–40. <https://doi.org/10.1016/j.geoderma.2014.03.021>
- McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1–2), 3–52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- McBratney, A., Minasny, B., MacMillan, R., and Carre, F. (2011). Digital Soil Mapping. In *Handbook of soil sciences: Properties and processes*. Boca Raton, FL: CRC Press.
- McKay, J., Grunwald, S., Shi, X., & Long, R. F. (2010). Evaluation of the transferability of a knowledge-based soil-landscape model. In *Digital Soil Mapping* (pp. 165-178). Springer, Dordrecht.
- Miller, B. A., Koszinski, S., Hierold, W., Rogasik, H., Schröder, B., Van Oost, K., ... Sommer, M. (2016). Towards mapping soil carbon landscapes: Issues of sampling scale and transferability. *Soil and Tillage Research*, 156, 194–208. <https://doi.org/10.1016/j.still.2015.07.004>

- Mitchell, G.A. (1998). Acting Director's Letter: 100 Years of Alaska Agriculture. *Agroborealis* 30 (1): 1. , Spring 1998. School of Agriculture and Land Resources Management Agricultural and Forestry Experiment Station, University of Alaska Fairbanks.
- Mishra, U., & Riley, W. J. (2012). Alaskan soil carbon stocks: spatial variability and dependence on environmental factors. *Biogeosciences*, 9(9), 3637–3645. <https://doi.org/10.5194/bg-9-3637-2012>
- Mishra, U., & Riley, W. J. (2014). Active-Layer Thickness across Alaska: Comparing Observation-Based Estimates with CMIP5 Earth System Model Predictions. *Soil Science Society of America Journal*, 78(3), 894. <https://doi.org/10.2136/sssaj2013.11.0484>
- Moonjun, R., Farshad, A., Shrestha, D. P., & Vaiphasa, C. (2010). Artificial Neural Network and Decision Tree in Predictive Soil Mapping of Hoi Num Rin Sub-Watershed, Thailand. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 151–164). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_13](https://doi.org/10.1007/978-90-481-8863-5_13).
- Moran, C. J., & Bui, E. N. (2002). Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science*, 16(6), 533-549.
- Mulder, V. L., de Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping — A review. *Geoderma*, 162(1–2), 1–19. <https://doi.org/10.1016/j.geoderma.2010.12.018>
- National Soil Survey Staff. SSURGO/STATSGO2 Structural Metadata and Documentation; NRCS Soils. Retrieved November 5, 2017. [https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2\\_053631](https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053631)
- Nauman, T. W., & Thompson, J. A. (2014). Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma*, 213, 385–399. <https://doi.org/10.1016/j.geoderma.2013.08.024>

- Nauman, T. W., Thompson, J. A., Odgers, N. P., & Libohova, Z. (2012). Fuzzy disaggregation of conventional soil maps using database knowledge extraction to produce soil property maps. In *Digital Soil Assessments and Beyond: Proceedings of the Fifth Global Workshop on Digital Soil Mapping* (pp. 203-207).
- Pain, C. F. (2005). Size does matter: relationships between image pixel size and landscape process scales. In *MODSIM, 2005, International Congress of Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand Inc* (pp. 1430-1436).
- Panda, S. K. (2014). High-Resolution Permafrost Modeling in Denali National Park and Preserve (Natural Resource Technical Report No. NPS/CAKN/NRTR—2014/858). National Park Service. Fort Collins, Colorado. Retrieved from <https://irma.nps.gov/App/Reference/Profile/2208990>
- Panda, S. K., Prakash, A., Solie, D. N., Romanovsky, V. E., & Jorgenson, M. T. (2010). Remote sensing and field-based mapping of permafrost distribution along the Alaska Highway corridor, interior Alaska. *Permafrost and Periglacial Processes*, 21(3), 271–281. <https://doi.org/10.1002/ppp.686>
- Pastick, N. J., Jorgenson, M. T., Wylie, B. K., Minsley, B. J., Ji, L., Walvoord, M. A., ... Rose, J. R. (2013). Extending Airborne Electromagnetic Surveys for Regional Active Layer and Permafrost Mapping with Remote Sensing and Ancillary Data, Yukon Flats Ecoregion, Central Alaska: Remote Sensing and Mapping of Permafrost and Active-layer Thickness. *Permafrost and Periglacial Processes*, 24(3), 184–199. <https://doi.org/10.1002/ppp.1775>
- Pastick, N. J., Rigge, M., Wylie, B. K., Jorgenson, M. T., Rose, J. R., Johnson, K. D., & Ji, L. (2014). Distribution and landscape controls of organic layer thickness and carbon within the Alaskan Yukon River Basin. *Geoderma*, 230–231, 79–94. <https://doi.org/10.1016/j.geoderma.2014.04.008>

- Ping, C. L., Bockheim, J. G., Kimble, J. M., Michaelson, G. J., & Walker, D. A. (1998). Characteristics of cryogenic soils along a latitudinal transect in arctic Alaska. *Journal of Geophysical Research: Atmospheres*, 103(D22), 28917–28928.  
<https://doi.org/10.1029/98JD02024>
- Ping, C. L., Clark, M. H., & Swanson, D. K. (2004). Cryosols in Alaska. *In* J.M. Kimble Ed. *Cryosols* (pp. 71-94). Springer, Berlin, Heidelberg.
- Ping, C.-L., Michaelson, G. J., Jorgenson, M. T., Kimble, J. M., Epstein, H., Romanovsky, V. E., & Walker, D. A. (2008a). High stocks of soil organic carbon in the North American Arctic region. *Nature Geoscience*, 1(9), 615–619. <https://doi.org/10.1038/ngeo284>
- Ping, C. L., Michaelson, G. J., Kimble, J. M., Romanovsky, V. E., Shur, Y. L., Swanson, D. K., & Walker, D. A. (2008b). Cryogenesis and soil formation along a bioclimate gradient in Arctic North America. *Journal of Geophysical Research*, 113(G3).  
<https://doi.org/10.1029/2008JG000744>
- Ping, C.-L., Clark, M. H., Kimble, J. M., Michaelson, G. J., Shur, Y., & Stiles, C. A. (2013). Sampling Protocols for Permafrost-Affected Soils. *Soil Horizons*, 54(1), 13.  
<https://doi.org/10.2136/sh12-09-0027>
- Qi, J., Kerr, Y., & Chehbouni, A. (1994). External factor consideration in vegetation index development. *Proc. of Physical Measurements and Signatures in Remote Sensing, ISPRS*, 723-730.
- R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104.  
<https://doi.org/10.1016/j.isprsjprs.2011.11.002>

- Roecker, S. M., & Thompson, J. A. (2010). Scale Effects on Terrain Attribute Calculation and Their Use as Environmental Covariates for Digital Soil Mapping. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 55–66). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_5](https://doi.org/10.1007/978-90-481-8863-5_5)
- Roecker, S. M., Howell, D. W., Haydu-Houdeshell, C. A., & Blinn, C. (2010). A Qualitative Comparison of Conventional Soil Survey and Digital Soil Mapping Approaches. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 369–384). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_29](https://doi.org/10.1007/978-90-481-8863-5_29)
- Roudier, P., Beaudette, D. E., & Hewitt, A. E. (2012). A conditioned Latin hypercube sampling algorithm incorporating operational constraints. *Digital Soil Assessments and Beyond*; CRC Press: Sydney, NSW, Australia, 227-231.
- Rozenstein, O., & Karnieli, A. (2011). Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. *Applied Geography*, 31(2), 533–544. <https://doi.org/10.1016/j.apgeog.2010.11.006>
- Scull, P., Franklin, J., & Chadwick, O. A. (2005). The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling*, 181(1), 1–15. <https://doi.org/10.1016/j.ecolmodel.2004.06.036>
- Sherwood, M. B. (1965). *Exploration of Alaska, 1865-1900*. Yale University Press, New Haven.
- Shi, X., Long, R., Dekett, R., & Philippe, J. (2009). Integrating Different Types of Knowledge for Digital Soil Mapping. *Soil Science Society of America Journal*, 73(5), 1682. <https://doi.org/10.2136/sssaj2007.0158>
- Shi, X., Zhu, A.-X., Burt, J. E., Qi, F., & Simonson, D. (2004). A Case-based Reasoning Approach to Fuzzy Soil Mapping. *Soil Science Society of America Journal*, 68(3), 885. <https://doi.org/10.2136/sssaj2004.8850>

- Soil Survey Staff. (1999). *Soil Taxonomy: A basic system of soil classification for making and interpreting soil surveys*. 2nd edition. Natural Resources Conservation Service. U.S. Department of Agriculture Handbook 436.
- Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. <http://websoilsurvey.nrcs.usda.gov>. Accessed February 2, 2018.
- Stepinski, T. F., & Jasiewicz, J. (2011). Geomorphons—a new approach to classification of landforms. *Proceedings of Geomorphometry 2011*, 109-112.
- Stum, A. K., Boettinger, J. L., White, M. A., & Ramsey, R. D. (2010). Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 179–190). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_15](https://doi.org/10.1007/978-90-481-8863-5_15)
- Thompson, J. A., Bell, J. C., & Butler, C. A. (2001). Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, 100(1–2), 67–89. [https://doi.org/10.1016/S0016-7061\(00\)00081-1](https://doi.org/10.1016/S0016-7061(00)00081-1)
- US Geological Survey (2016). *Landsat 8 Data Users Handbook*. USGS Publication LSDS-1574, Version 2.0.
- Van de Voorde, T., De Genst, W., & Canters, F. (2007). Improving pixel-based VHR land-cover classifications of urban areas with post-classification techniques. *Photogrammetric Engineering and Remote Sensing*, 73(9), 1017.
- Walker, D. A., Hamilton, T. D., Maier, H. A., Munger, C. A., & Reynolds, M. K. (2014). Glacial History and Long-Term Ecology in the Toolik Lake Region. In J. E. Hobbie & G. W. Kling (Eds.), *Alaska's Changing Arctic* (pp. 61–80). Oxford University Press. <https://doi.org/10.1093/acprof:osobl/9780199860401.003.0003>



- Wang, L., & Liu, H. (2006). An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International Journal of Geographical Information Science*, 20(2), 193–213.  
<https://doi.org/10.1080/13658810500433453>
- Washburn, A. L. (1973). *Periglacial Processes and Environments*. St. Martin's Press, New York.
- Webster, R., & Oliver, M. A. (2001). *Geostatistics for environmental scientists*. John Wiley & Sons.
- Wells, A., Macander, M., Jorgenson, M. T., Christopherson, T., Baird, B., Trainor, E., & Martyn, P. (2013, June). Ecological Land Classification, Soil Landscape Mapping, and Near Infrared (NIR) Spectroscopy of Soils in Lake Clark National Park and Preserve. *Alaska Park Science*, 12(1), 51–59.
- Wilson, F.H., Hults, C.P., Mull, C.G, and Karl, S.M, comps. (2015). *Geologic map of Alaska: U.S. Geological Survey Scientific Investigations Map 3340*, pamphlet 196 p., 2 sheets, scale 1:1,584,000. <http://dx.doi.org/10.3133/sim3340>
- Zhang, L., Sun, X., Wu, T., & Zhang, H. (2015). An Analysis of Shadow Effects on Spectral Vegetation Indexes Using a Ground-Based Imaging Spectrometer. *IEEE Geoscience and Remote Sensing Letters*, 12(11), 2188–2192. <https://doi.org/10.1109/LGRS.2015.2450218>
- Zhu, A. X. (2000). Mapping soil landscape as spatial continua: the neural network approach. *Water Resources Research*, 36(3), 663-677.
- Zhu, A.X., B. Hudson, J. Burt, K. Lubich, and D. Simonson. (2001). Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal* 65:1463-1472.

Table 2.1 Pedon locations and classifications used for modeling

Site ID	Series concept	Latitude (WGS84)	Longitude (WGS84)	Order	Suborder	Great group	Subgroup	Particle size class	Full taxonomic class
2016AK185003	ATIGUN	69.123723	-148.867309	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, dysic, Terric Hemistel
2016AK185004	ATIGUN	69.123532	-148.867423	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, dysic, Terric Hemistel
2016AK185008	ATIGUN	69.324078	-148.725413	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, dysic, Terric Hemistel
S2016AK185006	ATIGUN	68.981997	-148.841261	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, dysic, Terric Hemistel
2016AK185002	DALTON	69.124289	-148.870374	Gelisol	Turbel	Histoturbel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Histoturbel
2016AK185007	DALTON	69.323992	-148.727448	Gelisol	Turbel	Histoturbel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Histoturbel
2016AK185011	DALTON	68.98336	-148.845074	Gelisol	Turbel	Histoturbel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Histoturbel
2017AK185937	DALTON	68.333113	-149.349258	Gelisol	Turbel	Histoturbel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Histoturbel
2017AK185940	DALTON	68.397344	-149.323612	Gelisol	Turbel	Histoturbel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Histoturbel
2017AK185901	GALBRAITH	68.540278	-149.418115	Gelisol	Orthel	Aquorthel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Aquorthel
2017AK185930	GALBRAITH	68.454142	-149.497778	Gelisol	Orthel	Aquorthel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Aquorthel
93AK185003	GALBRAITH	69.15	-148.853889	Gelisol	Turbel	Aquiturbel	Typic	Fine-loamy	Fine-loamy, Typic Aquiturbel
93AK185006	GALBRAITH	68.62	-149.311389	Gelisol	Turbel	Aquiturbel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Aquiturbel
95AK185007	GALBRAITH	68.62	-149.616722	Gelisol	Turbel	Aquiturbel	Typic	Fine-loamy	Fine-loamy, Typic Aquiturbel
95AK185012	GALBRAITH	68.61	-149.307333	Gelisol	Orthel	Aquorthel	Typic	Fine-loamy	Fine-loamy, Typic Aquorthel
95AK185013	GALBRAITH	69.44	-148.666639	Gelisol	Turbel	Aquiturbel	Typic	Fine-loamy	Fine-loamy, Typic Aquiturbel
95AK185014	GALBRAITH	69.4	-148.797444	Gelisol	Turbel	Aquiturbel	Typic	Fine-silty	Fine-silty, Typic Aquiturbel
95AK185015	GALBRAITH	69.13	-148.58775	Gelisol	Orthel	Aquorthel	Typic	Fine-silty	Fine-silty, Typic Aquorthel
95AK185021	GALBRAITH	68.62	-149.2995	Gelisol	Turbel	Aquiturbel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Aquiturbel
98AK185001	GALBRAITH	70.163435	-148.454293	Gelisol	Turbel	Aquiturbel	Ruptic-Histic	Fine-loamy	Fine-loamy, Ruptic-Histic Aquiturbel
98AK185002	GALBRAITH	68.62	-149.609861	Gelisol	Turbel	Aquiturbel	Typic	Fine-loamy	Fine-loamy, Typic Aquiturbel
98AK185003	GALBRAITH	68.069722	-149.579806	Gelisol	Turbel	Aquiturbel	Typic	Coarse-silty	Coarse-silty, Typic Aquiturbel
CP06S 933	GALBRAITH	69.44	-148.629783	Gelisol	Turbel	Aquiturbel	Typic	Fine-loamy	Fine-loamy, Typic Aquiturbel
S01AK-185-002	GALBRAITH	69.67	-148.721306	Gelisol	Turbel	Aquiturbel	Ruptic-Histic	Coarse-silty	Coarse-silty, Ruptic-Histic Aquiturbel
S01AK-185-004	GALBRAITH	69.15	-148.848528	Gelisol	Turbel	Aquiturbel	Ruptic-Histic		Ruptic-Histic Aquiturbel
S1993AK185002	GALBRAITH	68.600281	-149.635834	Gelisol	Orthel	Aquorthel	Typic	Coarse-loamy	Coarse-loamy, Typic Aquorthel
S2003AK185002	GALBRAITH	69.395134	-148.736084	Gelisol	Turbel	Aquiturbel	Ruptic-Histic	Fine-loamy	Fine-loamy, Ruptic-Histic Aquiturbel
S2004AK185002	GALBRAITH	69.131111	-148.842499	Gelisol	Turbel	Aquiturbel	Typic	Fine-silty	Fine-silty, Typic Aquiturbel
S2004AK185006	GALBRAITH	68.069267	-149.580032	Gelisol	Turbel	Aquiturbel	Typic	Fine-silty	Fine-silty, Typic Aquiturbel
S2016AK185005	GALBRAITH	69.120518	-148.850563	Gelisol	Turbel	Aquiturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Aquiturbel
S2016AK185007	GALBRAITH	69.442939	-148.63244	Gelisol	Turbel	Aquiturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Aquiturbel
S2016AK185008	GALBRAITH	68.070579	-149.580228	Gelisol	Turbel	Aquiturbel	Glacic	Coarse-loamy	Coarse-loamy, Glacic Aquiturbel
S2016AK185009	GALBRAITH	69.115676	-148.852342	Gelisol	Turbel	Aquiturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Aquiturbel

Table 2.1 (cont.) Pedon locations and classifications used for modeling

Site ID	Series concept	Latitude (WGS84)	Longitude (WGS84)	Order	Suborder	Great group	Subgroup	Particle size class	Full taxonomic class
2017AK185929	GORGE	68.414673	149.310467	Inceptisol	Gelept	Haploglept	Aquic	Loamy-skeletal	Loamy-skeletal, calcareous, Aquic Haploglept
INTERP_12	Gravel	69.009437	148.815154	Gravel	Gravel	Gravel	Gravel	Gravel	Gravel
INTERP_13	Gravel	69.079826	148.737892	Gravel	Gravel	Gravel	Gravel	Gravel	Gravel
INTERP_4	Gravel	69.913398	148.725399	Gravel	Gravel	Gravel	Gravel	Gravel	Gravel
INTERP_6	Gravel	69.863745	148.732018	Gravel	Gravel	Gravel	Gravel	Gravel	Gravel
INTERP_7	Gravel	69.762433	148.681081	Gravel	Gravel	Gravel	Gravel	Gravel	Gravel
2016AK185009	ICECUT	68.983225	148.843347	Inceptisol	Gelept	Haploglept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haploglept
2016AK185012	ICECUT	68.811342	148.819804	Entisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
2016AK185013	ICECUT	68.811615	148.819214	Entisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
2016AK185015	ICECUT	68.896658	148.866987	Inceptisol	Gelept	Haploglept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haploglept
2016AK185017	ICECUT	68.894688	148.861233	Inceptisol	Gelept	Haploglept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haploglept
2016AK185018	ICECUT	68.983475	148.843825	Entisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
93AK185005	IMNAVAIT	68.63	149.582222	Gelisol	Orthel	Haplothel	Typic	Fine-loamy	Fine-loamy, Typic Haplothel
95AK185016	IMNAVAIT	69.06	148.745333	Gelisol	Turbel	Haploturbel	Typic	Sandy-skeletal	Sandy-skeletal, Typic Haploturbel
S01AK-185-003	IMNAVAIT	69.67	148.721222	Gelisol	Turbel	Haploturbel	Aquic	Coarse-loamy	Coarse-loamy over sandy or sandy-skeletal, Aquic Haploturbel
2017AK185912	ITKILLIK	68.716857	149.038373	Inceptisol	Gelept	Haploglept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haploglept
2017AK185914	ITKILLIK	68.64281	149.548857	Inceptisol	Gelept	Haploglept	Typic	Fragmental	Fragmental, Typic Haploglept
2017AK185922	ITKILLIK	68.22374	149.404132	Inceptisol	Gelept	Haploglept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haploglept
2017AK185935	ITKILLIK	68.548409	149.497071	Inceptisol	Gelept	Haploglept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haploglept
2017AK185909	IVISHAK	68.76495	148.902263	Entisol	Fluvent	Gelifluvent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelifluvent
2017AK185933	IVISHAK	68.199063	149.402728	Entisol	Fluvent	Gelifluvent	Aquic	Loamy-skeletal	Loamy-skeletal, Aquic Gelifluvent
S04AK-185-005	IVISHAK	67.969666	-149.77161	Entisol	Fluvent	Gelifluvent	Aquic	Loamy-skeletal	Loamy-skeletal, calcareous, Aquic Gelifluvent
2017AK185921	IVISHAK-Gravel	68.368862	-149.27141	Entisol	Fluvent	Gelifluvent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelifluvent
2017AK185932	IVISHAK-Gravel	68.20116	149.398089	Entisol	Fluvent	Gelifluvent	Aquic	Loamy-skeletal	Loamy-skeletal, Aquic Gelifluvent

Table 2.1 (cont.) Pedon locations and classifications used for modeling

Site ID	Series concept	Latitude (WGS84)	Longitude (WGS84)	Order	Suborder	Great group	Subgroup	Particle size class	Full taxonomic class
2017AK185903	KANAYUT-Rock Outcrop	68.53724	-149.400693	Entisol	Orthent	Gelorthent	Typic	Coarse-loamy	Loamy-skeletal, Typic Gelorthent
INTERP_17	KANAYUT-Rock Outcrop	68.443918	-149.299176	Gelisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
INTERP_18	KANAYUT-Rock Outcrop	68.444958	-149.299287	Entisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
INTERP_19	KANAYUT-Rock Outcrop	68.41439	-149.602087	Gelisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
INTERP_20	KANAYUT-Rock Outcrop	68.413628	-149.601838	Entisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
INTERP_21	KANAYUT-Rock Outcrop	68.306336	-149.539291	Entisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
INTERP_22	KANAYUT-Rock Outcrop	68.307425	-149.539096	Gelisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
INTERP_23	KANAYUT-Rock Outcrop	68.150115	-149.348768	Gelisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
INTERP_24	KANAYUT-Rock Outcrop	68.148918	-149.348832	Entisol	Orthent	Gelorthent	Typic	Loamy-skeletal	Loamy-skeletal, Typic Gelorthent
2016AK185016	KUPARUK	68.896341	-148.865877	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, euic, Terric Hemistel
2017AK185931	KUPARUK	68.470509	-149.490795	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, euic, Terric Hemistel
STRINGBOG1	KUPARUK	68.366651	-149.320709	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, euic, Terric Hemistel
STRINGBOG2	KUPARUK	68.851884	-148.833431	Gelisol	Histel	Hemistel	Terric	Loamy	Loamy, euic, Terric Hemistel
INTERP_1	Lake Water	70.212856	-148.461154	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water
INTERP_10	Lake Water	69.714665	-148.545019	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water
INTERP_15	Lake Water	68.632329	-149.605163	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water
INTERP_16	Lake Water	68.462314	-149.419419	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water
INTERP_2	Lake Water	70.207053	-148.177975	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water
INTERP_3	Lake Water	70.083298	-148.182188	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water
INTERP_9	Lake Water	69.746968	-148.508439	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water	Lake Water
2017AK185927	MOLAR	68.453474	-149.332971	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, calcareous, Typic Haplogelept
2017AK185938	MOLAR	68.41338	-149.313822	Inceptisol	Gelept	Haplogelept	Typic	Fragmental	Fragmental, calcareous, Typic Haplogelept
2017AK185939	MOLAR	68.410999	-149.319973	Inceptisol	Gelept	Haplogelept	Typic	Fragmental	Fragmental, calcareous, Typic Haplogelept
2017AK185928	MOLAR-Scree	68.44637	-149.355228	Inceptisol	Gelept	Haplogelept	Typic	Fragmental	Fragmental, calcareous, Typic Haplogelept
2017AK185916	MOUTONNE E	68.212635	-149.451824	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
2017AK185924	MOUTONNE E	68.491327	-149.387365	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
2017AK185934	MOUTONNE E	68.156798	-149.4414	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
2017AK185936	MOUTONNE E	68.357455	-149.323913	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept

Table 2.1 (cont.) Pedon locations and classifications used for modeling

Site ID	Series concept	Latitude (WGS84)	Longitude (WGS84)	Order	Suborder	Great group	Subgroup	Particle size class	Full taxonomic class
2017AK185900	MOUTONNEE-Scree	68.540497	-149.445937	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
2017AK185915	MOUTONNEE-Scree	68.212022	-149.43996	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
2017AK185923	MOUTONNEE-Scree	68.483976	-149.381076	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
95AK185017	MOUTONNEE-Scree	68.76	-149.4085	Inceptisol	Gelept	Dystrogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Dystrogelept
2017AK185917	PING	68.20932	-149.459773	Inceptisol	Gelept	Haplogelept	Aquic	Loamy-skeletal	Loamy-skeletal, Aquic Haplogelept
2017AK185926	PING	68.480157	-149.391234	Inceptisol	Gelept	Haplogelept	Aquic	Loamy-skeletal	Loamy-skeletal, Aquic Haplogelept
S07AK001-004	PING	68.62	-149.303111	Inceptisol	Gelept	Dystrogelept	Aquic	Loamy-skeletal	Loamy-skeletal, Aquic Dystrogelept
WET-GR	PING	68.281546	-149.379261	Inceptisol	Gelept	Haplogelept	Aquic	Loamy-skeletal	Loamy-skeletal, Aquic Haplogelept
S07AK001-002	Rock Outcrop-MOUTONNEE	68.13	-149.478111	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
2017AK185907	SAGAVINIRKTO K	68.867222	-148.844153	Entisol	Fluvent	Gelifluvent	Aquic	Coarse-loamy	Coarse-loamy, Aquic Gelifluvent
2016AK185001	SAGWON	69.124392	-148.870886	Gelisol	Orthel	Mollorthel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Mollorthel
2016AK185005	SAGWON	69.323662	-148.731122	Gelisol	Orthel	Mollorthel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Mollorthel
2016AK185006	SAGWON	69.323689	-148.729324	Gelisol	Orthel	Mollorthel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Mollorthel
2016AK185010	SAGWON	68.983569	-148.843724	Gelisol	Orthel	Mollorthel	Typic	Loamy-skeletal	Loamy-skeletal, Typic Mollorthel
S04AK-185-004	SAGWON	67.949772	-149.777077	Gelisol	Geloll	Haplogeloll	Typic	Sandy-skeletal	Sandy-skeletal, Typic Haplogeloll
INTERP_11	Stream Water	69.015589	-148.818073	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water
INTERP_14	Stream Water	68.907447	-148.826755	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water
INTERP_5	Stream Water	69.898978	-148.728628	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water
INTERP_8	Stream Water	69.770918	-148.653385	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water	Stream Water
2017AK185905	SWALE	68.918129	-148.881925	Entisol	Aquent	Gelaquent	Typic	Coarse-loamy	Coarse-loamy, Typic Gelaquent
2017AK185906	SWALE	68.925835	-148.881674	Gelisol	Histel	Hemistel	Typic	Histic	Euic, Typic Hemistel
CP11S 62	SWALE	68.609825	-149.3141	Gelisol	Histel	Sapristel	Fluvaquentic	Histic	Dysic, Fluvaquentic Sapristel
2017AK185902	TAPS	68.540877	-149.401993	Entisol	Orthent	Gelorthent	Typic	Fragmental	Fragmental, Typic Gelorthent
95AK185020	TAPS	68.51	-149.183639	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept
2017AK185919	TAPS-Scree	68.368113	-149.297317	Entisol	Orthent	Gelorthent	Typic	Fragmental	Fragmental, Typic Gelorthent
2017AK185925	TAPS-Scree	68.493247	-149.402299	Inceptisol	Gelept	Haplogelept	Typic	Loamy-skeletal	Loamy-skeletal, Typic Haplogelept

Table 2.1 (cont.) Pedon locations and classifications used for modeling

Site ID	Series concept	Latitude (WGS84)	Longitude (WGS84)	Order	Suborder	Great group	Subgroup	Particle size class	Full taxonomic class
2016AK185022	TOOLIK	69.120317	-148.851305	Gelisol	Turbel	Histoturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Histoturbel
2016AK185023	TOOLIK	69.120065	-148.851138	Gelisol	Turbel	Histoturbel	Glacic	Coarse-loamy	Coarse-loamy, Glacic Histoturbel
2017AK185904	TOOLIK	68.534088	-149.401759	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
2017AK185908	TOOLIK	68.851153	-148.840268	Gelisol	Turbel	Histoturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Histoturbel
2017AK185910	TOOLIK	68.645904	-149.407031	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
2017AK185911	TOOLIK	68.649261	-149.411316	Gelisol	Turbel	Histoturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Histoturbel
2017AK185913	TOOLIK	68.710517	-149.015047	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
2017AK185918	TOOLIK	68.207022	-149.410119	Gelisol	Turbel	Histoturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Histoturbel
2017AK185920	TOOLIK	68.368213	-149.280381	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
93AK185004	TOOLIK	68.63	-149.635833	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
95AK185005	TOOLIK	69.94	-148.808056	Gelisol	Turbel	Histoturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Histoturbel
95AK185006	TOOLIK	69.93	-148.8	Gelisol	Turbel	Histoturbel	Typic	Coarse-silty	Coarse-silty, Typic Histoturbel
96AK185001	TOOLIK	69.4	-148.785111	Gelisol	Turbel	Histoturbel	Ruptic	Fine-loamy	Fine-loamy, Ruptic Histoturbel
S01AK-185-005	TOOLIK	70.377064	-148.554721	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
S03AK-185-001	TOOLIK	68.63	-149.64495	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
S03AK-185-004	TOOLIK	69.981104	-148.696776	Gelisol	Orthel	Historthel	Typic	Coarse-loamy	Coarse-loamy, Typic Historthel
S2004AK185001	TOOLIK	69.604309	-148.646393	Gelisol	Turbel	Histoturbel	Typic	Fine-loamy	Fine-loamy, Typic Histoturbel
S2016AK185003	TOOLIK	69.117565	-148.851581	Gelisol	Turbel	Histoturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Histoturbel
S2016AK185004	TOOLIK	69.119599	-148.85112	Gelisol	Turbel	Histoturbel	Typic	Coarse-loamy	Coarse-loamy, Typic Histoturbel

Table 2.2 Environmental covariate layers included in each raster stack.

		Stacks							
		CATONLY	FEAT	GEO	GM	NOCAT	PCA-GM	PCA-GEO	PCA_NOCAT
Data Layers	Geomorphons	X			X		X		
	Morphometric Features	X	X						
	Geology	X		X				X	
	Maximum Curvature		X	X	X	X			
	Multiresolution Valley Bottom Flatness		X	X	X	X			
	Normalized Height		X	X	X	X			
	Potential Incoming Solar Radiation		X	X	X	X			
	Slope		X	X	X	X			
	Terrain Ruggedness Index (10 cell radius)		X	X	X	X			
	Terrain Ruggedness Index (25 cell radius)		X	X	X	X	X	X	X
	MSAVI2		X	X	X	X	X	X	X
	PCA Band 1						X	X	X
	PCA Band 2						X	X	X

Table 2.3 Results of visual evaluations by three soil scientists. Each mapset received a maximum of 60 points from each individual evaluation, and the results are sorted by evaluator mean rating.

	Map ratings (out of 60)				Mean rating	Taxonomic class	Resolution	Stack	Trees	OOB accuracy
MAPSET	Soil Scientist 1	Soil Scientist 2	Soil Scientist 3							
3	39	59	48		48.7	Great Group	10m	GEO	100	47.69
17	47	60	38		48.3	Series Concept	30m	FEAT	500	48.46
2	35	56	47		46	Great Group	10m	GM	100	48.46
18	46	50	42		46	Series Concept	30m	FEAT	100	46.92
20	41	52	41		44.7	Series Concept	30m	NOCA T	100	45.38
16	38	57	38		44.3	Series Concept	30m	NOCA T	250	49.23
14	35	57	38		43.3	Particle Size Class	30m	GEO	1000	61.24
8	26	59	42		42.3	Order	10m	GEO	100	74.62
19	38	51	38		42.3	Series Concept	30m	GEO	100	46.15
22	26	58	43		42.3	Suborder	10m	GM	1000	63.08
12	38	54	33		41.7	Particle Size Class	30m	FEAT	500	62.69
21	25	60	39		41.3	Suborder	10m	GEO	100	65.38
25	28	57	39		41.3	Suborder	10m	FEAT	100	60
5	37	51	35		41	Great Group	30m	GEO	1000	46.15
13	36	53	34		41	Particle Size Class	30m	NOCA T	1000	62.02
24	32	57	34		41	Suborder	30m	GM	100	60.77
15	35	56	31		40.7	Particle Size Class	30m	GEO	100	60.47
11	41	51	27		39.7	Particle Size Class	30m	NOCA T	500	63.57
1	29	59	28		38.7	Great Group	30m	GEO	500	49.23
23	29	53	34		38.7	Suborder	30m	FEAT	1000	62.31
10	21	58	26		35	Order	10m	PCA- GEO	250	73.85
6	22	52	29		34.3	Order	10m	PCA- GEO	100	76.15
4	32	35	26		31	Great Group	100m	FEAT	100	46.92
7	26	42	21		29.7	Order	100m	GM	750	75.38
9	26	39	24		29.7	Order	100m	GEO	250	74.62



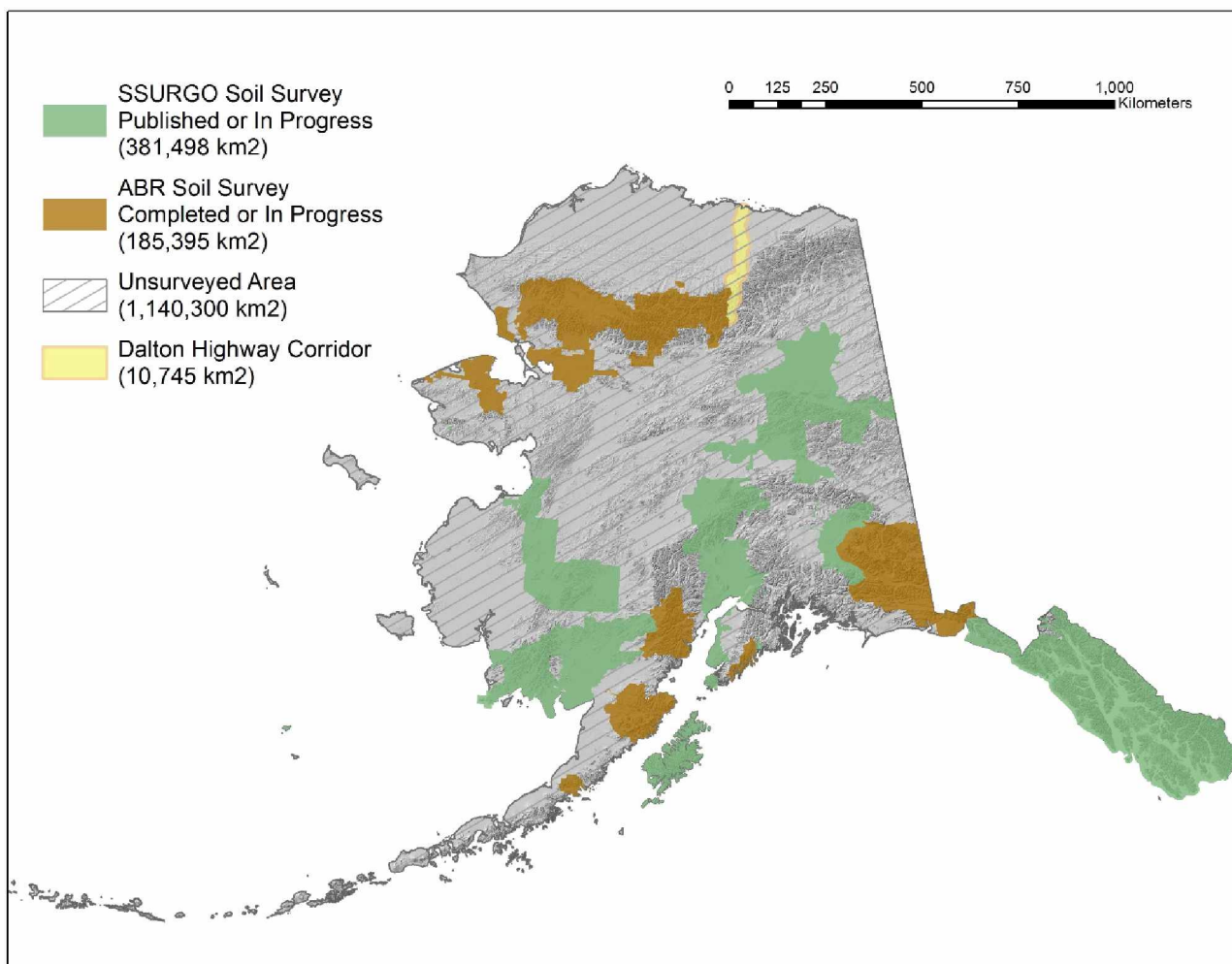


Figure 2.1 Status map showing SSURGO soil surveys, ABR soil surveys, and areas unsurveyed at scales finer than 1:500,000 in Alaska. The Dalton Highway corridor DSM research area is also highlighted.

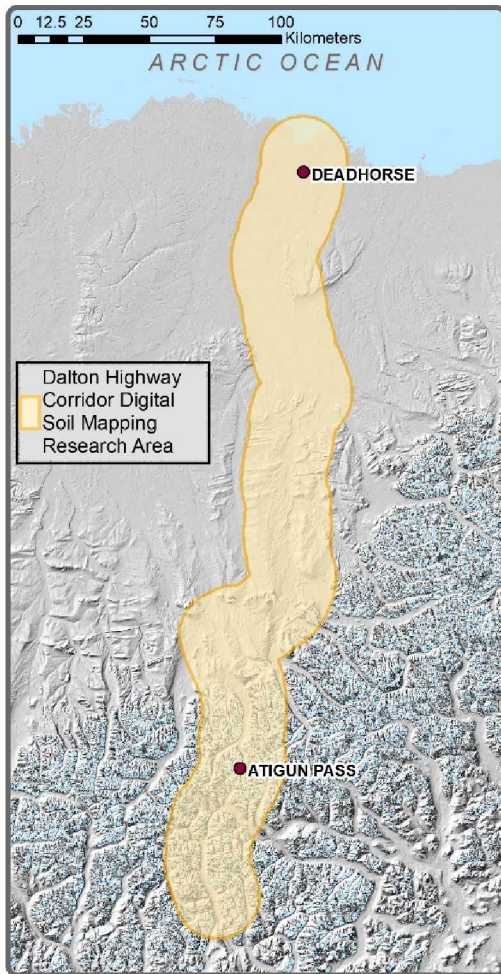


Figure 2.1 - Polygonal patterned ground in the Arctic Coastal Plain region. This area includes nearly level plains dissected by floodplain and terrace complexes of major rivers. Oblong thaw lakes are common. Soils are very wet, with accumulation of organic matter and ice-rich permafrost. Parent materials are generally cryoturbated fluvial deposits, with substantial micro-relief. Ponding occurs throughout the summer in low-centered polygons and in depressions overlying subterranean ice wedges.



Figure 2.2 - Rolling, glaciated slopes of the Arctic Foothills region. This region includes plains, kame and kettle complexes, floodplain and terrace complexes along major rivers, and occasional bedrock outcroppings and low mountains. Older glaciated surfaces have rounded slope shapes and are loess covered. Soils are commonly wet, with accumulation of organic matter and permafrost on nearly all slope positions. Younger glaciated surfaces have variable slope shapes and exposed surface fragments. Soils are generally rocky, dry, and lack permafrost on convex to linear positions.



Figure 2.3 - Typical landscape on warm slopes of the Brooks Range Mountains. The Dalton Highway and the Trans-Alaska Pipeline are visible in the photo as they follow the Atigun River valley. This subsection of the research area is entirely within the alpine life zone. Soils are generally very rocky and have little soil development. Colder and more protected positions have organic matter accumulations and permafrost despite having a high percentage of rock fragments. Rock outcrops are common on summits and shoulders, and scree slopes are common on both warm and cold backslope positions.

Figure 2.2 Detailed map of Dalton Highway corridor DSM research area, with landscape photos and descriptions of major physiographic regions.





Figure 2.3. Soil pit photos and descriptions of series concepts. a.) Photo of soil pit 2016AK185010 described as the SAGWON series concept (Loamy-skeletal, Typic Mollorthels). These soils are well-drained and are found at the convex edges of relict river terraces. The ICECUT series concept is found on similar landforms, but lacks mollic properties; b.) Photo of soil pit 2016AK185015 described as the ICECUT series concept (Loamy-skeletal, Typic Haplogelepts). These soils are well-drained and are found at the convex edges of relict river terraces. The SAGWON series concept is found on similar landforms, but has mollic properties.

Figure 2.3 (cont.) c.) Photo of soil pit 2017AK185916 described as the MOUTONNEE series concept (Loamy-skeletal, Typic Haploglepts). These soils are well-drained and are found on steep colluvial slopes. The MOLAR series concept is found on similar landforms but is calcareous; d.) Photo of soil pit 2017AK185928 described as the MOLAR series concept (Fragmental, Typic Haploglept). These soils are calcareous, well-drained, and are found on steep colluvial slopes. The MOUTONNEE series concept is found on similar landforms but is not calcareous; e.) Photo of soil pit 2017AK185919 described as the TAPS series concept (Fragmental, Typic Gelorthent). These soils are well-drained and are found on cold, steep colluvial slopes. The KANAYUT series concept is found on similar landforms but is shallow to bedrock. Both TAPS and KANAYUT commonly co-occur with bedrock outcroppings and scree slopes; f.) Photo of soil pit 2017AK185917 described as the PING series concept (Loamy-skeletal, Aquic Haploglepts). These soils are poorly-drained and are found in depressions and concavities on steep colluvial slopes. The GORGE series is found on similar landforms but is calcareous; g.) Photo of soil pit 2017AK185929 described as the GORGE series concept (Loamy-skeletal, Aquic Haploglept). These soils are calcareous, poorly-drained, and are found in depressions and concavities on steep colluvial slopes. The PING series concept is found on similar landforms but is not calcareous; h.) Photo of soil pit 2017AK185914 described as the ITKILLIK series concept (Loamy-skeletal, Typic Haploglepts). These soils are well-drained and are found on convex slopes of till and outwash deposits. The IMNAVAIT series concept is found on similar landforms but is coarse-loamy and commonly has permafrost within 1m; i.) Photo of soil pit 2017AK185918 described as the TOOLIK series concept (Coarse-loamy, Typic Histoturbels). These soils are poorly-drained and are found on linear to concave slopes of loess-capped till and outwash deposits. The DALTON series concept is found on similar landforms but is loamy-skeletal.

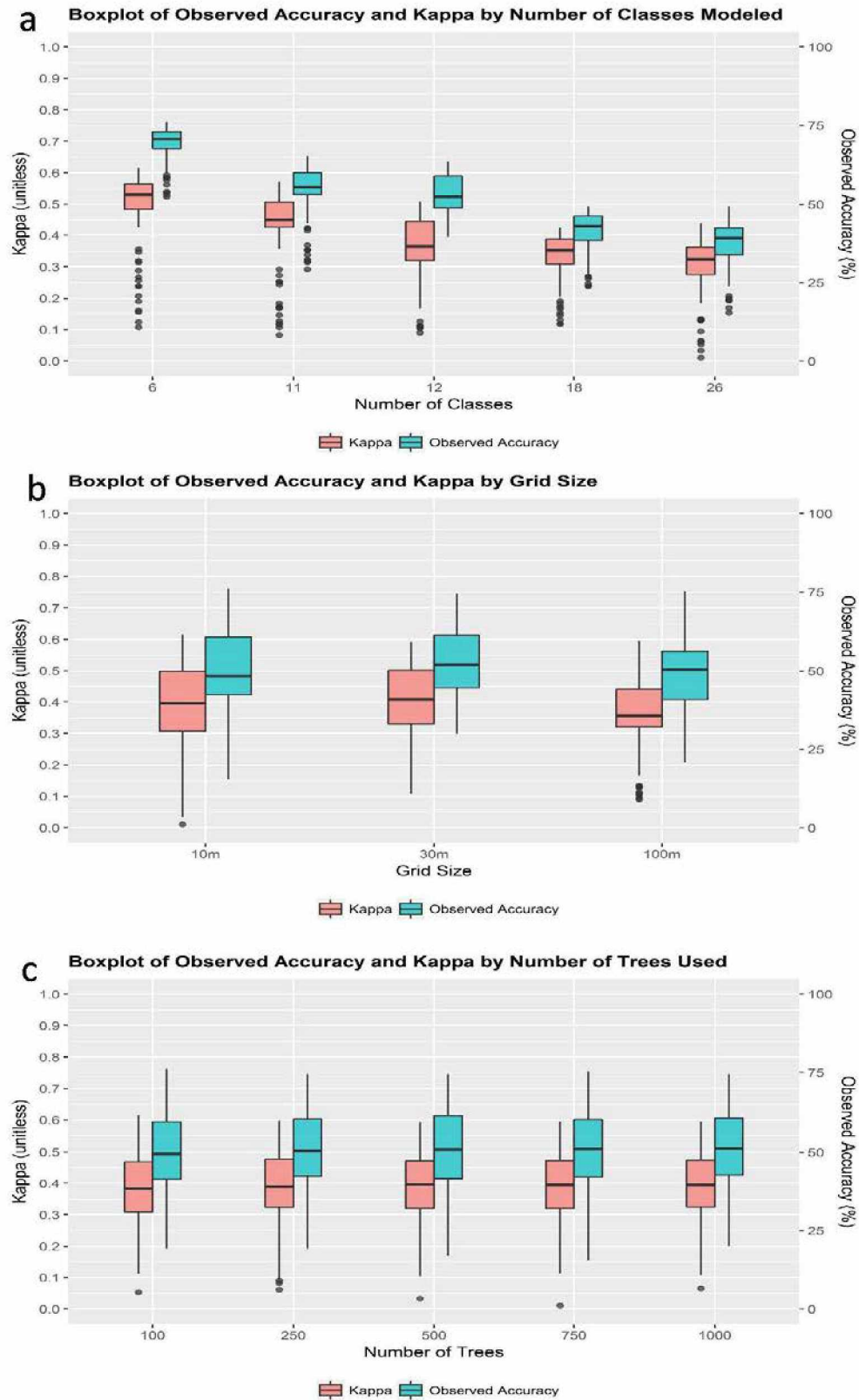


Figure 2.4 Boxplots of model accuracy. a.) Boxplot of model accuracy by number of classes modeled; b.) Boxplot of model accuracy by model grid size; c.) Boxplot of model accuracy by number of trees used.

## 2.10 Appendix A

Sample rubric for visual evaluation of map results.

DIGITAL SOIL MAPPING PILOT STUDY						
DALTON HIGHWAY CORRIDOR						
QUALITATIVE MAP RATING RUBRIC						
<b>MAP LEGEND LABEL:</b>						
		poor	fair	good	very good	excellent
BROOKS RANGE MTNS		1	2	3	4	5
	summits, shoulders, and ridgelines					
	backslopes					
	upland drainageways					
	lowland drainageways (incl. Floodplains, terraces, gravel bars)					
ARCTIC FOOTHILLS						
	hilltops, summits (convexities)					
	backslopes					
	upland drainageways					
	lowland drainageways (incl. Floodplains, terraces, gravel bars)					
ARCTIC COASTAL PLAIN						
	rises (convexities)					
	talfs, dips, swales (concavities)					
	lowland drainageways (incl. Floodplains, terraces, gravel bars)					
	lakes					

## 2.11 Appendix B

List of terrain attributes and SAGA geoprocessing modules.

<b><i>SAGA Geoprocessing</i></b>		
<b>Terrain attribute</b>	<b>SAGA library</b>	<b>SAGA module</b>
Morphometric Features	Morphometry	Morphometric Features
Maximum Curvature	Morphometry	Slope, Aspect, Curvature
Multiresolution Valley Bottom Flatness	Morphometry	Multiresolution Index of Valley Bottom Flatness
Normalized Height	Morphometry	Relative Heights and Slope Positions
Potential Incoming Solar Radiation	Lighting, Visibility	Potential Incoming Solar Radiation
Slope	Morphometry	Slope, Aspect, Curvature
Terrain Ruggedness Index	Morphometry	Terrain Ruggedness Index

## **CHAPTER 3: DIGITAL SOIL MAPPING OF PERMAFROST SOILS ALONG THE DALTON HIGHWAY CORRIDOR, ALASKA: A PILOT STUDY USING A RANDOM FOREST CLASSIFIER<sup>3</sup>**

### **3.1 Abstract**

This study utilizes a random forest classifier to predict multiple soil taxonomic classes from a basic collection of environmental covariates generated using high resolution (10m) satellite images and sparsely sampled pedon data. Covariates included maximum curvature, multiresolution valley bottom flatness, normalized height, potential incoming solar radiation, slope, terrain ruggedness index, and modified soil and vegetation index. Five tiers of soil taxonomic units are predicted: Order, Suborder, Great Group, Series Concept, and Particle Size Class. Model outputs are compared quantitatively via estimated out-of-bag accuracy. Estimated out-of-bag accuracy ranged from ~45% to ~75%, with results improving when fewer classes were modeled. We suggest future research into optimized sampling to ensure an adequate distribution of samples across the feature space. Overall, digital soil mapping with random forest classifiers appears to be a promising method for completing the soil survey of Alaska.

---

<sup>3</sup> Paul, J. et al., 2018, Multiresolution Digital Soil Mapping of Permafrost Soils Using a Random Forest Classifier, (prepared for submission to Journal of Permafrost and Periglacial Processes)



### 3.2 Introduction

Digital mapping of continuous or classified soil properties in Alaska has grown along with worldwide interest in climate change. Climate modeling in the remote, sparsely sampled Arctic and subarctic regions of Alaska demands rapid production of datasets representing soil information. The parameters of soil organic carbon (SOC) stocks and active layer thickness (ALT) have been mapped extensively using remote sensing techniques to meet these needs, but mapping of soil types and taxonomic units has not received equal attention (Deluca and Boisvenue, 2012; Hugelius, 2012; Hugelius et al., 2014; Jafarov et al., 2012; Mishra and Riley, 2012, 2014; Panda et al., 2010; Panda, 2014; Pastick et al., 2013, 2014; Ping et al., 2008a). USDA-NRCS has completed statewide soils mapping at the Digital General Soil Map of the United States standard (STATSGO2, 1:63,360 and coarser scales), but has only mapped approximately 20% of the state at the finer-resolution Soil Survey Geographic Database standard (SSURGO, 1:63,360 and finer scales) (National Soil Survey Staff, 2017; Soil Survey Staff, 2018) (Figure 2.1).

In this research, we meet the Soil Survey Geographic Database (SSURGO) standard of mapping taxonomic soil classes at scales of less than or equal to 1:63,360 by applying a random forest classification method (National Soil Survey Staff, 2017). Using common environmental covariates at high resolution (10m), this study presents a baseline modeling accuracy that can be expected when using freely available data layers and limited distribution of sampling points. The overall accuracy of map results is assessed using both point-based ground truth data. With a sampling density of 106 direct soil observations for a mapping area of 12,088 km<sup>2</sup> (1 observation per ~114 km<sup>2</sup>), this research will serve as an appropriate pilot study for digital soil mapping in remote areas of Alaska with limited pedon data.

#### 3.2.1 Study Area

The research area is an approximately 10 mile corridor on either side of the Dalton Highway, extending from Atigun Pass in the Brooks Range to the terminus of the highway at

Deadhorse, Alaska (Figure 2.2). The majority of the study area is within the zone of continuous permafrost, where over 90% of the earth's surface contains materials at or below 0 C for two or more consecutive years (Washburn, 1973). Three physiographic regions are contained within the study area boundary (Figure 2.2), with parent materials including rocky colluvium, glacial till and outwash, and fluvial sediments (Gallant et al., 1995; Hamilton, 2003; Huryn and Hobbie, 2012; Ping et al., 1998, 2008b, 2013; Walker et al., 2014).

### **3.3 Data Acquisition and Preprocessing**

#### **3.3.1 Soils Data**

Historical field pedon data, pedon descriptions from field work specific to this study, and remotely sensed points were aggregated into a single point data file for use in model training and accuracy assessment. All pedons were described to NRCS standards and most can be classified to the subgroup or family level. The total number of pedons in the dataset was 106, with an additional 24 points classified indirectly via remote sensing (Table 2.1). These additional points were in miscellaneous areas such as water bodies, gravel bars, and rock outcrops where the regolith is typically not classified as a soil but components are traditionally still included in soil map units as miscellaneous land types.

Point data attributes included multiple levels of soil taxonomic classifications ranging from soil order down to particle-size class (Soil Survey Staff, 1999). As soil series have not yet been established for this region, unique "series concept" names and their corresponding taxonomic units were created specifically for this research and will be proposed to soil survey staff (Figure 2.3). In this mapping context, these series concept names function similarly to complexes or associations of soil series, but individual units are not as narrowly defined with regard to Soil Taxonomy. Each pedon was correlated to the series concept of closest fit based on landform, parent material, and basic taxonomic unit. As a result, pedon data labeled with a unique series concept name may have

multiple taxonomic classifications associated with it. The “series concept” can be considered as a cluster of closely related soil types with similar interpretive properties.

Some points in the dataset are classified with multiple, concatenated series concept names and miscellaneous areas. These points represent complexes of soil and miscellaneous areas that occur at the sub-pixel scale (for example, shallow soils co-occurring with rock outcrops on a mountain summit). These points occur mostly in the mountainous regions of the mapping areas where "pure" pixels of a given soil type are uncommon.

### **3.3.2 Environmental Covariates**

The collection of environmental covariates used in modeling includes commonly used terrain attributes along with a modified soil vegetation index (MSAVI2) layer derived from mosaicked Landsat 8 imagery (Table 2.2). The following terrain attributes were derived from the Interferometric Synthetic Aperture Radar (IFSAR) Digital Terrain Model (DTM): maximum curvature, multiresolution valley bottom flatness (MRVBF), normalized height, potential incoming solar radiation (PISR), slope, and terrain ruggedness index (TRI). See Paul et al., (2018) for a detailed description of the data layers and the preprocessing methods used, and additional data layers that were considered but ultimately not included in this research.

## **3.4 Methods**

### **3.4.1 Random Forest Classification**

A random forest classification method was chosen here both for its use in previous soil and ecological mapping research (Chan and Paelinckx, 2008; Hengl et al., 2015; Rodriguez-Galiano, et al., 2012; Roecker, et al., 2010; Stum et al., 2010), and for the option of using all available data for accuracy assessment (described below). The pedon dataset was also quite imbalanced with regard to classes (some classes were represented by less than 3 points), and the random forest method has

been shown to perform well with imbalanced data when compared to other machine learning classifiers (Khoshgoftaar, 2007).

The random forest machine learning algorithm takes the ensemble approach to decision tree modeling and includes at least two layers of randomness in each tree. Model training data (in this case, soil pedon classifications and their spatial locations) are bootstrapped for training each individual tree, while the nodes of each individual tree are also limited by using a small, random selection of layers within the input layer stack (Breiman, 2001; Liaw and Wiener, 2002). Final predicted classes are decided by receiving the majority of votes from all trees.

Processing was completed with the "raster" and "randomForest" packages for R, the latter built around the original Breiman (2001) algorithm (Hijmans, 2016; Liaw and Wiener, 2002, 2008; R Core Team, 2016). Using previous research as a guide, random forests were built with 100 trees (Paul et al., 2018).

### **3.4.2 Accuracy Assessment**

Because each individual tree is built using a random selection of ground-truth soil class data (a bootstrap sample), the random forest method eliminates the need for setting aside a percentage of our already limited data points for validation (Breiman, 2001).

In contrast to the historical shortcomings of soil class mapping as reviewed by Brevik et al. (2016), this modern classification method allows for an internal quantitative accuracy assessment by aggregating the "out-of-bag" (OOB) error rates from each individual tree (Liaw and Wiener, 2002). This estimate of OOB error is built in to the randomForest package and accompanies each modeled map output, and was used in this study to evaluate model results.

### **3.5 Results and Discussion**

#### **3.5.1 Estimated OOB Accuracy and Kappa by Taxonomic Class**

Estimated OOB accuracies and kappa values varied by taxonomic class predicted, ranging from 69% and 0.51 (respectively) for soil Order to 44% and 0.37 for soil Series Concept (Table 3.1).

#### **3.5.2 Importance Table**

The randomForest package provides an optional importance table output with mean decrease in accuracy (MDA) values computed for each layer used in the random forest classification (Table 3.2). Though both Gini importance and MDA are computed, Gini importance is shown to be more biased towards predictor variables with many categories (Strobl et al., 2007). Since categorical data was included in some stacks, only MDA after permutation was computed here. To calculate MDA for a data layer, values in that layer are permuted (rearranged randomly) between all data points for each model run; the average difference in accuracy between these permuted models and the original models is the MDA for that layer. This provides the user with a relative measure of layer importance and an estimate of model performance if a layer was not used (Archer and Kimes, 2008).

### **3.6 Discussion**

#### **3.6.1 Importance of Environmental Covariates**

MDA values identified a similar hierarchy of layer importance for each taxonomic class modeled. Normalized height and PISR were generally ranked as the least important layers, and MSAVI2, TRI (10 cell radius), and MRVBF were generally ranked as most important.

The environmental covariates chosen for this study are frequently used digital soil mapping research with the exception of the PISR layer, which was included here to identify warm and cold slope positions commonly recognized as soil forming factors in the Arctic and sub-Arctic.

However, it appears that including PISR does not provide as much information to the model as would be expected. This is surprising given the emphasis on solar radiation in most state factor models of soil development in permafrost zones. There is possibly a redundancy with the MSAVI2 layer, as vegetation type and plant productivity are known to vary with regard to slope and exposure to sunlight. With the exception of the immediate road corridor, the vegetation in the study area is almost entirely undisturbed by human land use and is unaffected by wildfire, making it a suitable proxy for solar radiation. Though the correlation of predictor variables is shown to have a positive effect on importance measures, low MDA values for PISR and high values for MSAVI2 lead us to conclude that these layers are not strongly correlated in this research (Smith et al., 2011; Strobl et al., 2008).

In a similar way, including both the normalized height layer and MRVBF may have been redundant as conceptually these layers should be the inverse of one another. Flat valley bottoms will be represented by high values in MRVBF and low values in a normalized height layer, and vice versa. It is worth noting that these layer importance measures are calculated across the study area as a whole; we could expect a different hierarchy of layer importance if the study area was stratified by physiographic region, with PISR likely becoming more important in high relief areas.

There is the option to use principle component analysis (PCA) to reduce redundancy in the data, but previous research in this study area shows that replacing environmental covariates with their PCA bands has a negative impact on accuracy, as does including categorical data layers representing geology and/or landforms (Paul et al., 2018). This is surprising as most soil mapping approaches recognize soil-landform and soil-parent material relationships as major factors in developing soil components and map units. Either these categorical layers were poor representations of their respective phenomena, or continuous terrain attributes and vegetation indices are simply more powerful predictors of soil classes.

### 3.6.2 Covariate Resolution

While it may seem intuitive to use a high resolution dataset in this research, interpretation of these results is confounded by the fact that the study area includes both high and low relief landscapes and accuracy is computed for the study area as a whole. Thompson et al., (2001) and Pain et al., (2005) have shown that high resolution DEMs may have the greatest positive effect on map accuracy when used in high relief areas, while Cavazzi et al., (2013) actually noted a decrease in random forest model accuracy using high resolution DEM in low relief areas.

In addition to model resolution, Maynard and Johnson (2014) found window size to have a larger impact on soil property predictions than DEM resolution when using a high resolution (1-5 m) dataset. Roecker and Thompson (2010) determined that terrain attributes that compute surface curvatures are most sensitive to window size, with optimal windows determined with reference to the size of local landforms present in the mapping area. Window sizes were not investigated systematically in this study, though it is a promising direction for future research. The inclusion of local mean filters alongside raw covariate data by Moran and Bui (2002) also provides an example of using neighborhood analysis instead of varying DEM resolution. The challenge in classifying a large area that varies from high to low relief is to find resolutions and window sizes that provide adequate detail where needed without overanalyzing the terrain in other areas.

The idea of physiographic stratification was not overlooked in this study; the intent was to test if the random forest method could perform a meaningful landscape stratification from the data provided. Our visual inspection (detailed below) shows that soil classes generally appear in appropriate geomorphic positions. Most geomorphic misclassification is observed where soils exclusively located on mountain slopes are predicted in lower hillslope areas, or where soils exclusively located on floodplains appear in upland areas. In the future, incorporating local mean filters or outright geomorphic stratification might improve or maintain accuracy while using coarser resolution (>10m) DTMs.

### **3.6.3 Classification Schema**

Accuracies of model results were closely related to the level of taxonomy used to classify the input point data. This is likely due to increasingly specific levels of taxonomy having more unique soil types. Therefore, it is appropriate to consider the number of unique classes predicted in the random forest model as the independent variable rather than taxonomic class. As the number of unique classes to predict increases, accuracy appears to decrease (Table 3.1). The Particle Size Class taxonomic group, which is an ancillary classification not directly related to any other soil taxonomic units outside of Series Concept, appears to support this relationship between accuracy and number of unique classes. The question of which classification scheme is best will no doubt depend on the final use of the map. As a general guideline, map producers using similar methods should expect accuracies of less than 50% when attempting to predict 15 or more classes.

### **3.6.4 Pedon Data**

It is unclear how the model accuracy was affected by point data distribution or quantity. Since this study used a mixture of legacy pedon data and landform-based transect data, input point data was both unevenly distributed in geographical space and of very low density.

Random forest algorithms rely on spatial association of soil properties or classes to environmental covariates throughout the feature space, with each point evaluated individually from adjacent points. As such, random forests are not necessarily sensitive to the geographic distribution of point data, but are sensitive to point data distribution within the feature space. As detailed in Bui et al., (2006), prediction of soil classes in unobserved areas can fail when sampling design does not include representative areas for the entire feature space. In contrast, spatial interpolation methods (kriging or cokriging) rely on the spatial autocorrelation of observations and therefore perform best with high-density sampling in geographic space. Predictions are also typically not extended beyond the sampling extent. Miller et al., (2016) found that multiple logistic regression performs better than kriging or cokriging when modeling soil properties outside of the sampling extent, suggesting that



even among geostatistical methods spatial autocorrelation may be a poor choice when mapping remote areas.

Brungard and Boettinger (2010) used a conditioned Latin Hypercube Sampling (cLHS) method to determine an optimal sampling size of 200-300 points to adequately represent the feature space in a  $\sim 300\text{km}^2$  study area. While the approach is admirable, this level of sampling density (1 to 1.5 observations per  $\text{km}^2$ ) is logistically implausible in almost all unmapped areas of Alaska. Liess (2015) raises the issue of sampling design in DSM where using proper geostatistical sampling methods is not possible, and suggests an optimization process that starts with determining the accessible zones within the mapping area and dividing the feature space within those accessible zones. In a similar fashion, Roudier et al., (2012) incorporate a “cost” layer into the cLHS optimization. A hybrid approach including legacy data alongside an optimized sampling design would be required for this specific project where some data exists but may not be entirely representative.

It is likely that point data distribution may affect model results more than point data quantity. For example, when using a boosted classification tree, Grinand (2008) did not find an increase in classification accuracy of external areas by increasing internal sampling density. More methods to optimize sampling design in a machine learning context should be explored in the future, with special attention paid to remote areas with difficult access. The difficulty of traveling across tundra on foot also contributed to clustering of point data locations, with soil scientists choosing to transect regions where a maximum of landform variability could be observed with a minimum of effort. This method is efficient in the field, but often causes two unique soil classes to be observed in adjacent pixels, or even within one pixel when using coarse resolution data. This poor spacing of observations and possible inappropriate description methods highlight the need for an accepted soil sampling protocol for patterned ground features, like those proposed by Ping et al., (2013).

### **3.6.5 Quantitative Accuracy**

Beyond training the model, point data was also used to determine quantitative model accuracies. When predicting soil properties, Bishop et al., (2015) show that using point-based ground truth data for validation often presents a worst-case scenario for map accuracy. Mean values derived from block supports are shown to have higher accuracy ratings, with supports based on grids or polygons surrounding point-based data. This method has not yet been incorporated into the random forest accuracy assessment and would likely require a separate validation dataset, which is limiting in the context of mapping remote areas with logistical constraints and small input datasets. However, expert knowledge combined with other remote sensing strategies could be used to delineate validation areas and compare expected soil classes with model results.

More research is needed to determine the proper validation method for regional-scale soil maps and come to a consensus on acceptable accuracy levels for various uses of soil maps. As discussed by Baveye and Laba (2015), we must also ask what degree of heterogeneity needs to be conveyed to the end user, and at what confidence, when considering the interpretive value of soil class or soil property maps.

### **3.6.6 Recommendations for Post-Processing**

Though it is not traditionally a taxonomic class in and of itself, particle size class (PSC) was predicted with relatively high accuracy, having only 9 unique classes. Combining predicted PSC with another map output via a raster calculator function may provide additional interpretive value without altering the classification methodology. Individual soil Subgroup prefixes were also predicted with a very high accuracy, though the group had only 7 unique classes and ~80% of the input data was in the class of "Typic". The resulting map of almost entirely one class was not very useful, and these results are not included above for that reason. However, one could theoretically predict the PSC, Subgroup prefix, and Great Group separately and combine the outputs into a higher taxonomic class (e.g. taxon above family). It would be difficult to determine the accuracy of

such a product, and some class combinations may be impossible taxonomic units. The result would likely appear closer to a traditional SSURGO level component legend and would potentially increase map utility or aid in comparisons against traditional soil survey products.

Obvious mapping errors could be reduced by reclassifying pixels within basic physiographic stratifications (for instance, pixels within the Arctic Coastal Plain that are predicted as components sampled exclusively in the Brooks Range Mountains). This might be appropriate for the most egregious errors, but if landscape stratification is to be performed at all one might suggest modeling each strata separately from the beginning rather than in post-processing. A more appropriate workflow might be to reclassify pixels as complexes of multiple components using conditional reasoning. When classes are often observed or predicted closely together in geographic space, the pixels could be reclassified as a complex of the two classes to reduce the "confetti" effect of modeling two soils that share similar landforms and landform positions (e.g. Historthels and Histoturbels in the Arctic Coastal Plain). Reclassification could be done when pixels of two specified classes are adjacent, or when pixels are within some distance of each other.

In land cover mapping, "noisy" model outputs are commonly processed through boundary cleaning, majority filtering, or other workflows involving expert knowledge and/or ancillary reference data (Rozenstein and Karnieli, 2011; Van de Voorde, et al., 2007). When compared to continuous value rasters, classified outputs have fewer tools available as the processing is often conditional rather than arithmetic or statistical. Replacing single pixels or small clusters of pixels with neighboring classes can greatly enhance the visual appeal of the map and is a crucial step before converting the raster model to SSURGO-style polygons.

Finally, it is important to note the accuracy assessment performed by the random forest algorithm is valid for the raw model only. Ideally, a separate accuracy assessment would be performed for the finished, post-processed model. Combining classes into complexes or otherwise altering the original class structure would require a more complicated investigation of which validation points are misclassified. Again, workflows that rely on conditional statements may be

most appropriate. When comparing to raw random forest accuracies, a k-fold cross-validation method may be the most simple and straightforward assessment of post-processed model accuracy in this case.

### **3.7 Conclusions**

Overall, this research suggests that random forest modeling is a promising method for digital soil mapping in the sparsely sampled regions of Alaska. Results shown here are a marked improvement from the currently available STATSGO2 dataset due to the finer scale model output and the increased number of data points used to populate the model. Modeling with a simple stack of continuous environmental covariates provides accurate soil maps at taxonomic levels of great group and higher. Including a solar radiation layer may not significantly impact prediction of soil classes, even though solar radiation is often present in state factor models of soil development in permafrost zones.

Additional input is needed to determine the most reasonable number of classes required to provide adequate interpretive values for natural resource management on public and private lands. As with all soil surveys, stakeholders will need to discuss the scope of the project and determine the level of detail that will meet their needs. Though it is clear that map accuracy will decrease when more classes are modeled, maps can easily be produced at multiple taxonomic levels using this method, with almost no additional processing time. This could potentially allow end users a suite of maps to choose from depending on their needs, or contribute to a multiresolution digital soil mapping experience.

Optimization workflows for future soil sampling in remote regions should be pursued, and should ideally include legacy data wherever possible. Future soil sampling should be carried out with modeling resolution in mind to avoid sampling clusters, with additional efforts to consistently describe patterned ground and other periglacial features to improve model training.

### 3.8 Acknowledgments

This research was supported by the USDA Natural Resources Conservation Service grant (Award # 68-7482-15-531) and the UADA NIFA program. We would like to thank colleagues Stephanie Schmit, Eric Geisler, and Matt Ferderbar who provided soils expertise that greatly assisted the research. We also thank Dr. Zamir Libohova (USDA-NRCS) and Dr. Jordi Cristobal (University of Alaska, Fairbanks) for comments that greatly improved the manuscript.

### 3.9 References

- Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4), 2249–2260.  
<https://doi.org/10.1016/j.csda.2007.08.015>
- Baveye, P. C., & Laba, M. (2015). Moving away from the geostatistical lamppost: Why, where, and how does the spatial heterogeneity of soils matter? *Ecological Modelling*, 298, 24–38.  
<https://doi.org/10.1016/j.ecolmodel.2014.03.018>
- Bishop, T. F. A., Horta, A., & Karunaratne, S. B. (2015). Validation of digital soil maps at different spatial supports. *Geoderma*, 241–242, 238–249.  
<https://doi.org/10.1016/j.geoderma.2014.11.026>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.  
<https://doi.org/10.1023/A:1010933404324>
- Brevik, E. C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., & Jordán, A. (2016). Soil mapping, classification, and pedologic modeling: History and future directions. *Geoderma*, 264, 256–274. <https://doi.org/10.1016/j.geoderma.2015.05.017>

- Brungard, C. W., & Boettinger, J. L. (2010). Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 67–75). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_6](https://doi.org/10.1007/978-90-481-8863-5_6)
- Bui, E. N., Simon, D., Schoknecht, N., & Payne, A. (2006). Chapter 15 Adequate Prior Sampling is Everything: Lessons from the Ord River Basin, Australia. In *Developments in Soil Science* (Vol. 31, pp. 193–608). Elsevier. [https://doi.org/10.1016/S0166-2481\(06\)31015-X](https://doi.org/10.1016/S0166-2481(06)31015-X)
- Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., & Fealy, R. (2013). Are fine resolution digital elevation models always the best choice in digital soil mapping? *Geoderma*, 195–196, 111–121. <https://doi.org/10.1016/j.geoderma.2012.11.020>
- Chan, J. C.-W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, 112(6), 2999–3011. <https://doi.org/10.1016/j.rse.2008.02.011>
- Deluca, T. H., & Boisvenue, C. (2012). Boreal forest soil carbon: distribution, function and modelling. *Forestry*, 85(2), 161–184. <https://doi.org/10.1093/forestry/cps003>
- Gallant, A. L., Binnian, E. F., Omernik, J. M., & Shasby, M. B. (1995). *Ecoregions of Alaska* (USGS Numbered Series No. 1567) (p. 78). U.S. Geological Survey. Retrieved from <http://pubs.er.usgs.gov/publication/pp1567>
- Grinand, C., Arrouays, D., Laroche, B., & Martin, M. P. (2008). Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, 143(1–2), 180–190. <https://doi.org/10.1016/j.geoderma.2007.11.004>
- Hamilton, T. D. (2003). *Glacial geology of the Toolik Lake and upper Kuparuk River regions*. Institute of Arctic Biology. University of Alaska Fairbanks.

- Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., ... Tondoh, J. E. (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *PLOS ONE*, 10(6), e0125814. <https://doi.org/10.1371/journal.pone.0125814>
- Hijmans, R. (2016). raster: Geographic Data Analysis and Modeling. <https://CRAN.R-project.org/package=raster>
- Hugelius, G. (2012). Northern Circumpolar Soil Carbon Database. ECDS Environment Climate Data Sweden. <https://doi.org/10.5879/ecds/000000001>
- Hugelius, G., Strauss, J., Zubrzycki, S., Harden, J. W., Schuur, E. A. G., Ping, C.-L., ... Kuhry, P. (2014). Estimated stocks of circumpolar permafrost carbon with quantified uncertainty ranges and identified data gaps. *Biogeosciences*, 11(23), 6573–6593. <https://doi.org/10.5194/bg-11-6573-2014>
- Huryn, A. D., & Hobbie, J. E. (2012). Land of extremes: a natural history of the Arctic North Slope of Alaska. Fairbanks, Alaska: University of Alaska Press.
- Jafarov, E. E., Marchenko, S. S., & Romanovsky, V. E. (2012). Numerical modeling of permafrost dynamics in Alaska using a high spatial resolution dataset. *The Cryosphere*, 6(3), 613–624. <https://doi.org/10.5194/tc-6-613-2012>
- Khoshgoftaar, T. M., Golawala, M., & Hulse, J. V. (2007). An Empirical Study of Learning from Imbalanced Data Using Random Forest (pp. 310–317). *IEEE*. <https://doi.org/10.1109/ICTAI.2007.46>
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18–22.
- Liaw, A., & Wiener, M. (2008). randomForest: Breiman and Cutler's Random Forests for Classification and Regression. <http://CRAN.R-project.org/package=randomForest>

- Liess, M. (2015). Sampling for regression-based digital soil mapping: Closing the gap between statistical desires and operational applicability. *Spatial Statistics*, 13, 106–122.  
<https://doi.org/10.1016/j.spasta.2015.06.002>
- Maynard, J. J., & Johnson, M. G. (2014). Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: Effects of grid resolution vs. neighborhood extent. *Geoderma*, 230–231, 29–40. <https://doi.org/10.1016/j.geoderma.2014.03.021>
- Miller, B. A., Koszinski, S., Hierold, W., Rogasik, H., Schröder, B., Van Oost, K., ... Sommer, M. (2016). Towards mapping soil carbon landscapes: Issues of sampling scale and transferability. *Soil and Tillage Research*, 156, 194–208.  
<https://doi.org/10.1016/j.still.2015.07.004>
- Mishra, U., & Riley, W. J. (2012). Alaskan soil carbon stocks: spatial variability and dependence on environmental factors. *Biogeosciences*, 9(9), 3637–3645. <https://doi.org/10.5194/bg-9-3637-2012>
- Mishra, U., & Riley, W. J. (2014). Active-Layer Thickness across Alaska: Comparing Observation-Based Estimates with CMIP5 Earth System Model Predictions. *Soil Science Society of America Journal*, 78(3), 894. <https://doi.org/10.2136/sssaj2013.11.0484>
- Moran, C. J., & Bui, E. N. (2002). Spatial data mining for enhanced soil map modelling. *International Journal of Geographical Information Science*, 16(6), 533-549.
- National Soil Survey Staff. SSURGO/STATSGO2 Structural Metadata and Documentation; NRCS Soils. Retrieved November 5, 2017.  
[https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2\\_053631](https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053631)
- Pain, C. F. (2005). Size does matter: relationships between image pixel size and landscape process scales. In MODSIM, 2005, International Congress of Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand Inc (pp. 1430-1436).



- Panda, S. K. (2014). High-Resolution Permafrost Modeling in Denali National Park and Preserve (Natural Resource Technical Report No. NPS/CAKN/NRTR—2014/858). National Park Service. Fort Collins, Colorado. Retrieved from <https://irma.nps.gov/App/Reference/Profile/2208990>
- Panda, S. K., Prakash, A., Solie, D. N., Romanovsky, V. E., & Jorgenson, M. T. (2010). Remote sensing and field-based mapping of permafrost distribution along the Alaska Highway corridor, interior Alaska. *Permafrost and Periglacial Processes*, 21(3), 271–281. <https://doi.org/10.1002/ppp.686>
- Pastick, N. J., Jorgenson, M. T., Wylie, B. K., Minsley, B. J., Ji, L., Walvoord, M. A., ... Rose, J. R. (2013). Extending Airborne Electromagnetic Surveys for Regional Active Layer and Permafrost Mapping with Remote Sensing and Ancillary Data, Yukon Flats Ecoregion, Central Alaska: Remote Sensing and Mapping of Permafrost and Active-layer Thickness. *Permafrost and Periglacial Processes*, 24(3), 184–199. <https://doi.org/10.1002/ppp.1775>
- Pastick, N. J., Rigge, M., Wylie, B. K., Jorgenson, M. T., Rose, J. R., Johnson, K. D., & Ji, L. (2014). Distribution and landscape controls of organic layer thickness and carbon within the Alaskan Yukon River Basin. *Geoderma*, 230–231, 79–94. <https://doi.org/10.1016/j.geoderma.2014.04.008>
- Paul, J. D. (2018). Multiresolution Digital Soil Mapping of Permafrost Soils Using a Random Forest Classifier. Manuscript in preparation.
- Ping, C. L., Bockheim, J. G., Kimble, J. M., Michaelson, G. J., & Walker, D. A. (1998). Characteristics of cryogenic soils along a latitudinal transect in arctic Alaska. *Journal of Geophysical Research: Atmospheres*, 103(D22), 28917–28928. <https://doi.org/10.1029/98JD02024>
- Ping, C.-L., Michaelson, G. J., Jorgenson, M. T., Kimble, J. M., Epstein, H., Romanovsky, V. E., & Walker, D. A. (2008a). High stocks of soil organic carbon in the North American Arctic region. *Nature Geoscience*, 1(9), 615–619. <https://doi.org/10.1038/ngeo284>

- Ping, C. L., Michaelson, G. J., Kimble, J. M., Romanovsky, V. E., Shur, Y. L., Swanson, D. K., & Walker, D. A. (2008b). Cryogenesis and soil formation along a bioclimate gradient in Arctic North America. *Journal of Geophysical Research*, 113(G3).  
<https://doi.org/10.1029/2008JG000744>
- Ping, C.-L., Clark, M. H., Kimble, J. M., Michaelson, G. J., Shur, Y., & Stiles, C. A. (2013). Sampling Protocols for Permafrost-Affected Soils. *Soil Horizons*, 54(1), 13.  
<https://doi.org/10.2136/sh12-09-0027>
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104.  
<https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Roecker, S. M., & Thompson, J. A. (2010). Scale Effects on Terrain Attribute Calculation and Their Use as Environmental Covariates for Digital Soil Mapping. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 55–66). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_5](https://doi.org/10.1007/978-90-481-8863-5_5)
- Roecker, S. M., Howell, D. W., Haydu-Houdeshell, C. A., & Blinn, C. (2010). A Qualitative Comparison of Conventional Soil Survey and Digital Soil Mapping Approaches. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 369–384). Dordrecht: Springer Netherlands.  
[https://doi.org/10.1007/978-90-481-8863-5\\_29](https://doi.org/10.1007/978-90-481-8863-5_29)
- Roudier, P., Beaudette, D. E., & Hewitt, A. E. (2012). A conditioned Latin hypercube sampling algorithm incorporating operational constraints. *Digital Soil Assessments and Beyond*; CRC Press: Sydney, NSW, Australia, 227-231.

- Rozenstein, O., & Karnieli, A. (2011). Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. *Applied Geography*, 31(2), 533–544. <https://doi.org/10.1016/j.apgeog.2010.11.006>
- Smith, S. J., Ellis, N., & Pitcher, C. R. (2011). Conditional variable importance in R package extendedForest. R vignette. <http://gradientforest.r-forge.r-project.org/Conditional-importance.pdf>. Accessed March 12, 2018.
- Soil Survey Staff. 1999. Soil Taxonomy: A basic system of soil classification for making and interpreting soil surveys. 2nd edition. Natural Resources Conservation Service. U.S. Department of Agriculture Handbook 436.
- Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. <http://websoilsurvey.nrcs.usda.gov>. Accessed February 2, 2018.
- Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1), 25.
- Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1), 307.
- Stum, A. K., Boettinger, J. L., White, M. A., & Ramsey, R. D. (2010). Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah. In J. L. Boettinger, D. W. Howell, A. C. Moore, A. E. Hartemink, & S. Kienast-Brown (Eds.), *Digital Soil Mapping* (pp. 179–190). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-90-481-8863-5\\_15](https://doi.org/10.1007/978-90-481-8863-5_15)
- Thompson, J. A., Bell, J. C., & Butler, C. A. (2001). Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*, 100(1–2), 67–89. [https://doi.org/10.1016/S0016-7061\(00\)00081-1](https://doi.org/10.1016/S0016-7061(00)00081-1)
- Van de Voorde, T., De Genst, W., & Canters, F. (2007). Improving pixel-based VHR land-cover classifications of urban areas with post-classification techniques. *Photogrammetric Engineering and Remote Sensing*, 73(9), 1017.

Walker, D. A., Hamilton, T. D., Maier, H. A., Munger, C. A., & Raynolds, M. K. (2014). Glacial History and Long-Term Ecology in the Toolik Lake Region. In J. E. Hobbie & G. W. Kling (Eds.), *Alaska's Changing Arctic* (pp. 61–80). Oxford University Press.

<https://doi.org/10.1093/acprof:osobl/9780199860401.003.0003>

Washburn, A. L. (1973). *Periglacial Processes and Environments*. St. Martin's Press, New York.

Table 3.1 Estimated OOB Accuracy and Kappa by taxonomic level modeled.

<b>Taxonomic Level</b>	<b>Unique Classes</b>	<b>Estimated OOB Accuracy (%)</b>	<b>Kappa</b>
Order	6	69.23	0.51
Suborder	11	60	0.51
Particle Size Class	12	58.91	0.45
Great Group	18	46.92	0.4
Series Concept	26	43.85	0.37

Table 3.2 Mean Decrease in Accuracy (MDA) values for each environmental covariate and taxonomic level modeled.

	<b>Taxonomic Level</b>				
	<b>Order</b>	<b>Suborder</b>	<b>Great Group</b>	<b>Series Concept</b>	<b>Particle Size Class</b>
<b>Environmental Covariate Layer</b>					
Potential Incoming Solar Radiation	2.69	2.19	2.12	1.84	1.06
Normalized Height	0.94	3.24	2.85	2.48	3.35
Maximum Curvature	5.91	4.27	4.12	5.96	2.04
MSAVI2	8.45	5.7	5.31	7.15	5.57
Slope	4.45	5.8	4.64	5.6	5.54
Terrain Ruggedness Index (25 cell)	4.83	6.42	5.83	2.03	3.86
Terrain Ruggedness Index (10 cell)	6.94	8.15	6.2	6.44	7.22
Multiresolution Valley Bottom Flatness	6.69	8.72	7.4	6.75	5.86

## CHAPTER 4: CONCLUSION

In this research, the use of random forest machine learning algorithms was applied to digital soil mapping, with particular attention paid to the modeling parameters of resolution, data layers used, and number of trees built. More specifically, digital soil mapping with random forests was applied in a remote area in Alaska using limited environmental data layers and a set of sparsely sampled soils data comparable to what would be used in a production soil survey (approximately 1 observation per 114 km<sup>2</sup>). This research provides useful background information and a baseline for future digital soil mapping efforts, and should minimize questions about whether the correct modeling parameters were chosen for a particular project.

In chapter one, combinations of environmental input data were tested at multiple resolutions and five tiers of soil taxonomic units were predicted. Based on the digital soil mapping methodology investigated in chapter one, a pilot study was conducted in a remote area along the Dalton Highway in northern Alaska in chapter two. The main findings of both chapters were as follows:

- Using a random forest classifier with common environmental covariates, digital soil mappers can expect out-of-bag accuracy ratings of approximately 45% to 75% when modeling between 5 and 30 soil classes with sparsely sampled pedon data.
- Model accuracy appears to decrease with increasing number of classes when using this method, meaning that digital soil maps can be expected to lose accuracy at finer taxonomic resolutions.
- Including categorical environmental data such as landforms and geologic units does not noticeably improve random forest classification of soil types at any taxonomic level, nor does building the forest with more than 100 trees.
- Performing principal component analysis (PCA) on the environmental covariates and replacing these covariates with the first two PCA bands in the random forest model

consistently decreases model accuracy.

- Importance tables show that using a potential incoming solar radiation layer in the model did not improve model performance and was not as strong a predictor of high latitude soil types as is commonly expected.
- Qualitative visual evaluations of map results by soil scientists showed a marked preference for maps with more soil classes, even though these maps had the lowest quantitative accuracy. Conversely, maps depicting fewer soil classes scored the lowest in visual evaluations, but actually had the highest quantitative accuracy.

Model runs at 10m and 30m resolution performed comparably, with 100m resolution performing ~5-10% worse in most cases. These results are a marked improvement from the currently available STATSGO2 spatial dataset due to the finer scale model output and the increased number of data points used to populate the model. However, it should be noted that the maps are not directly comparable due to issues of scale and format; STATSGO2 is in vector format with aggregated soil components grouped into mapunits, while this research presents a digital soil map in raster format with individual soil components made spatially explicit.

Results suggested that random forest model performance was limited to out-of-bag accuracy rates of approximately 75% when modeling less than 10 soil classes, and approximately 45% when modeling 25 to 30 soil classes. Generally, results improved when fewer classes were modeled. Digital soil maps created via random forests can therefore be expected to lose accuracy at finer taxonomic resolutions. More research is needed to determine the most reasonable number of classes required to provide adequate interpretive values for natural resource management on public and private lands. As with all soil surveys, stakeholders will need to discuss the project and determine the level of detail that will meet their needs. Accuracy and taxonomic resolution should be considered within the design and scoping phase of a soil survey if using digital soil mapping methods, as providing an accuracy assessment is one of the benefits of digital soil mapping when



compared to traditional soil survey. Sampling design should reflect the intended accuracy and taxonomic resolution of the soil survey in order to avoid investigating fine taxonomic differences in soil types that are not able to be predicted spatially with high accuracy. Optimization workflows for future soil sampling in remote regions should be pursued, and should ideally include legacy data wherever possible. Future soil sampling should be carried out with modeling resolution in mind to avoid sampling clusters occurring within one pixel or adjacent pixels. Additional efforts to consistently describe patterned ground and other periglacial features may improve model training, especially where patterned ground cycles occurring at sub-pixel scale.

At all resolutions tested, using a small number of trees (100 or less) on a simple stack of continuous environmental covariates provided accurate soil maps at taxonomic levels of great group and higher. Building over 100 trees did not appear to improve model performance and was computationally expensive. Replacing environmental covariates with PCA bands also did not improve model performance. These pre-processing steps are considered to be unimportant to this research and I would not suggest their use in future digital soil mapping projects in northern Alaska. The addition of categorical environmental input data (e.g. landforms and geology) did not substantially increase map accuracy and should be considered unnecessary in this area, especially as the inclusion of these layers required additional software and involved substantial human processing time. In areas where soil types are very strongly correlated with landforms or geologic units, inclusion of layers representing these environmental covariates may deserve consideration provided these layers are spatially accurate. After analyzing importance table outputs from the random forest model, it appears that including a solar radiation layer also did not significantly impact prediction of soil classes, even though solar radiation is almost always present in state factor models of soil development in permafrost zones. Preprocessing of this layer was also computationally expensive and may be unnecessary in this area, provided that a suitable proxy (e.g., a vegetation index) is used in the model.

Visual evaluations by soil scientists showed highly unexpected results, with qualitative map ratings generally favoring maps with low quantitative accuracy. Though maps predicting Series Concept performed worst in accuracy values derived from point data, their high performance in visual evaluation shows that soils must have been modeled realistically with respect to landforms and landform positions. This apparent disagreement between model accuracy and visual ratings suggests that OOB estimates of accuracy should not be the sole method of evaluating DSM products. This speaks to the nature of soil survey as a process where tacit knowledge and local field experience are crucial in both the creation and evaluation of soil maps. This paradigm presents many challenges to DSM approaches in remote, sparsely sampled regions where there may be few people (if any) with the level of expert knowledge commonly attained during traditional soil surveys. There is still considerable opportunity to use expert knowledge combined with other remote sensing strategies for rigorous accuracy assessments. However, we can expect this process to be difficult if digital soil maps are presented in raster form and/or on a component scale (as in this research), since most soil scientists are familiar with vector-based aggregated map products. As stated above, point-based accuracy assessments may not be the best choice for DSM, and the author recommends incorporating more expert knowledge and other alternative methods of model assessment into future DSM research.

For broader applications beyond the usual scope of soil surveys, global climate modelers could also benefit from a wider application of random forest digital soil mapping throughout the circumpolar north using coarse resolution outputs and a small number of classes representing basic soil taxonomic groups. As always, the greatest challenge in such a large and remote region will be collecting representative pedon data. As previously stated, any new sampling conducted in support of a continental- or circumpolar-scale digital soil map should be optimized for the feature space used in the model, requiring extensive geospatial research, data acquisition, and preprocessing before sampling design is even considered.